THE AMERICAN MATHEMATICAL MONTHLY

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

Volume 103, Number 1 January 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

Richard BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

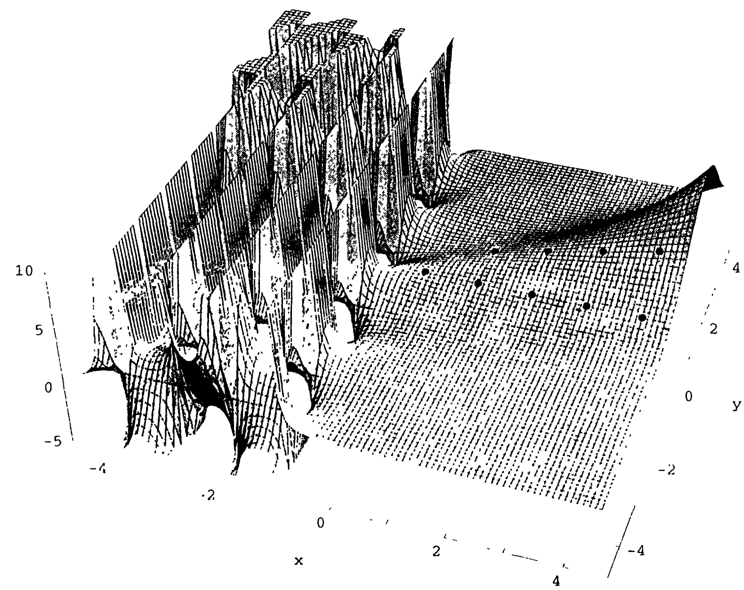
Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

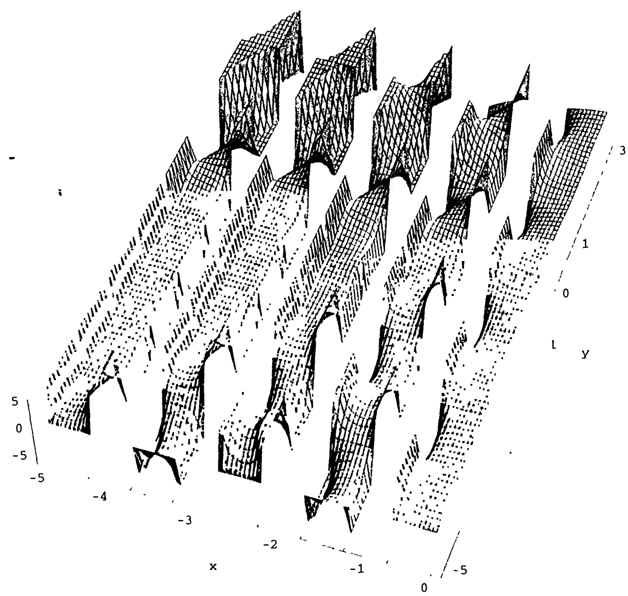
The Binomial Coefficient Function

David Fowler

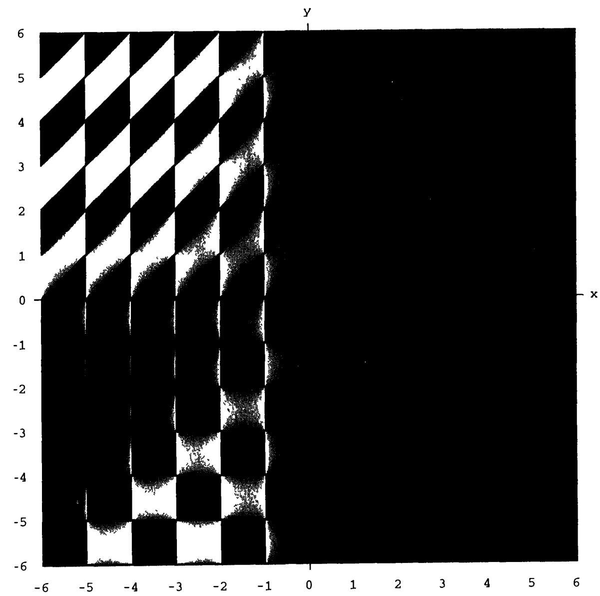
1. SOME PLOTS. The surface defined by the binomial function $C(x, y) = x!/y!(x-y)!$, with 'Pascal's ridge' rising on the right, looks like:



and an exploded view, looking into the 'Avenues of Manhattan', like:



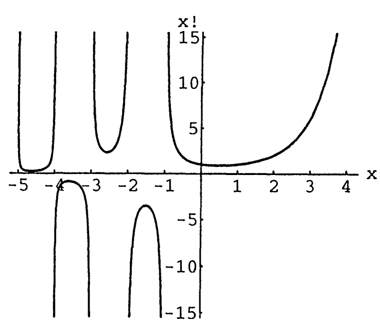
These plots cannot however represent the fine detail near to the lines of singularities $x = -1, -2, -3, \ldots,$ where for the most part the edges of the truncated surfaces appear misleadingly as ruler-straight. More information can be read off from the following 'density plot', in which height is represented by a gray scale from white $(-\infty)$ to black $(+\infty)$:

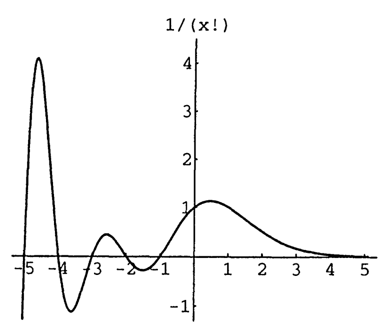


The scale here can be judged by looking at the section along $y = 1$ where, we shall see, $C(x, 1) = x$. The dominant mid-gray denotes numbers close to zero, and the horizontal and diagonal lines are the zeros of the function.

2. A GENERAL DESCRIPTION. The factorial function can be defined by $x! = \Gamma(x+1) = \int_0^\infty t^x e^{-t}\, dt$ for $x > -1$, extended uniquely to negative non-integral x by defining $x! = (x+n)!/(x+n)(x+n-1)\ldots(x+1)$ for any integer n such that $(x+n) > -1$. It is infinitely differentiable except at the negative integers, and the graphs of it and its reciprocal look like:

x! and 1/(x!)





The binomial function C is thus defined and smooth for $(x, y) \in \mathbf{R}^2 \setminus \{x = -1, -2, -3, \ldots\}$, and the singularities of $x!$ at $-1, -2, -3, \ldots$ give rise to the very complicated behavior of C for $x < -1 + \eta$, $\eta > 0$ (hereafter $0 < \eta < 1$, h is a real number, and n & m are positive integers). These lines of singularities of C may surprise and shock at first, but they are easy to explain: C inherits most of the basic identities satisfied by the binomial coefficients $\binom{n}{m}$—perhaps all that make sense in this wider context—so $C(x-1, y) = ((x-y)/x)C(x, y)$, from one of the most elementary properties of binomial coefficients. Setting $x = 0$, we see that C must have a singularity at $(-1, y)$ unless $C(0, y) = 0$ (so y is a non-zero integer) or $y = 0$; and the resulting line of singularities will then propagate backwards to $x = -2, -3, \ldots$. By examining sections of this surface, we see further that:

- For $0 \leq y \leq x$, it is Pascal's triangle interpolated to a steeply rising ridge.
- For $\{0 \leq x \leq y\} \cup \{-1 + \eta \leq x \leq 0, y \geq 0\}$, it is gently undulating, zero on the lines $y - x = 1, 2, 3, \ldots$, and becomes close to zero more than a unit away from the boundary lines.
- For $x < -1, y > 0$, away from the boundary, it quickly becomes very large, alternately positive and negative inside the parallelograms with corners $(-n, m+1), (-n, m), (-n-1, m-1), (-n-1, m)$, passing steeply through zero on the lines $y - x = 1, 2, 3, \ldots$, and going off to $\pm\infty$ on the lines $x = -1, -2, -3, \ldots$.
- For $y < 0$, it is zero on $y = -1, -2, -3, \ldots, x \neq -1, -2, -3, \ldots$, and very close to zero in the quadrant $x > -1, y < 0$ more than a unit away from the boundary. In the octant $0 > x > y$, its behavior is similar to that on $x < 0, y > 0$, but here based on the squares of the integer grid: zero on $y = -1, -2, -3, \ldots$, and infinite on $x = -1, -2, -3, \ldots$. And in the octant $-1 > y > x + 1$, away from the boundary, it becomes close to zero except near to the lines $x = -1, -2, -3, \ldots$, where it passes steeply through $\pm\infty$, always in the same sense: as x decreases from $-n$ to $-(n+1)$, it increases very rapidly from $-\infty$ to near 0, stays near 0, and then again increases rapidly to $+\infty$.
- C is very badly behaved near the lines $x = -1, -2, -3, \ldots$, although directional limits exist at the integer points on these lines. (Set $x = -n + t, y = \pm m + \lambda t$ in C and manipulate to get a function of t, well-behaved at 0, multiplied by $(-1+t)!/(-1+(1-\lambda)t)!$. Then this latter expression can be written $(t!/t)((1-\lambda)t/((1-\lambda)t)!)$, so it tends to $(1-\lambda)$ as t tends to 0.) For example, the directional limit at $(-1, 0)$ is 1 along the x-axis, 0 along the diagonal $y = x + 1$, and 2 along the diagonal $y = -x - 1$. In particular, the horizontal directional limits that appear in the horizontal sections $y = \pm m$ below give the classical binomial coefficients $\binom{x}{\pm m}$ for $x = -1, -2, -3, \ldots$ though, to repeat, $C(-n, m)$ is not defined. The example of $\binom{n}{m} = \binom{n}{n-m}$ but $\binom{-n}{m} \neq \binom{-n}{-n-m}$ is instructive: since $C(x, y) = C(x, x-y), x \neq -1, -2, -3, \ldots$, the vertical section $C(h, y)$, $h \neq -1, -2, -3, \ldots$, is symmetric about $y = \frac{1}{2}h$ because $C(h, \frac{1}{2}h + (k - \frac{1}{2}h)) = C(h, \frac{1}{2}h - (k - \frac{1}{2}h))$. Now take $h = nx, k = \pm mx$, and look along the lines through the origin: the two directional limits $\lim_{x \to -1} C(nx, \pm mx)$ and $\lim_{x \to -1} C(nx, (n \mp m)x)$ *are* equal. One could generalise the standard binomial coefficients to include an extra argument, the slope at which the directional limit is to be taken, and thus extend such standard identities to negative arguments and perhaps find new ones.

3. SOME SECTIONS OF *C*. In all formulae in this section, $x \neq -1, -2, -3, \ldots$ unless these values can be filled in by continuity; $0 < \eta < 1$; h denotes a real number; and n & m are positive integers, though occasionally, as in the expression

$\pm m$ above, we also silently include $m = 0$. If any formula contains a term $h!$ where h is negative, we can replace it with $(h + n)!/(h + n)(h + n - 1)\dots(h + 1)$ for any n such that $(h + n) \geq 0$. This will give a unique result since $(h + n + 1)! = (h + n + 1)(h + n)!$. One important instance, used repetitively below, is that $1/(-n)! = (-n + 1)\dots -1.0/0! = 0$.

A lot of the detail of the descriptions, and more, is illustrated in the graphs of sections of the surface, which are arranged as far as possible to correspond to the ways they cut the surface, even when this means they have to be rotated through 90°. Take careful note of their scales, because they vary from 10^{-2} to 10^5 and have been chosen to give the most detail in each individual case.

Vertical sections. We use the basic and elegant identity $\pi y/y!(-y)! = \sin \pi y$.

$x = 0$: $C(0, y) = 0!/y!(-y)! = \sin \pi y/\pi y$. In particular, this is 1 when $y = 0$, and 0 when $y = \pm m$.

$$x = n\text{:}\quad C(n, y) = \frac{n!}{y!(n - y)!}$$

$$= \begin{cases} \text{the classical binomial coefficient if } y = m, 0 \leq m \leq n \\ 0 \text{ if } y = \pm m \text{ outside this range} \\ \dfrac{n!}{(n - y)(n - 1 - y)\dots(1 - y)} \dfrac{\sin \pi y}{\pi y} \text{ in general.} \end{cases}$$

This quickly gets very small for large $|y|$, because of the polynomial in the denominator of degree $n + 1$ in y.

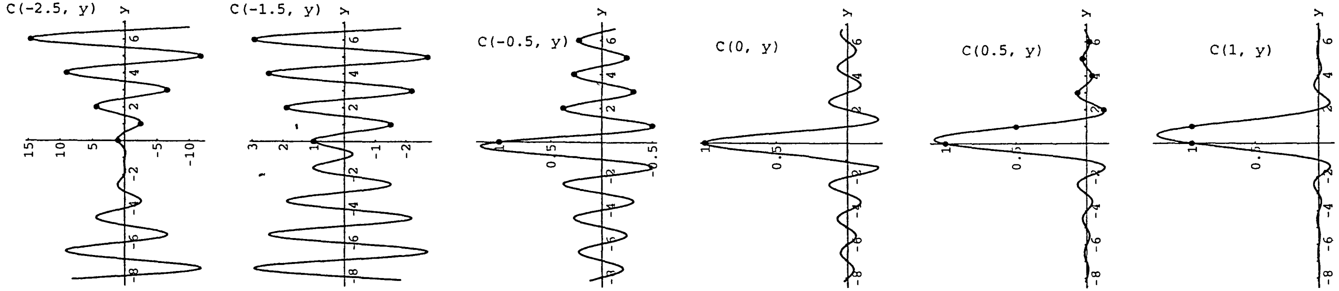
We can fit these special cases within the general context, as follows:

$x = -h, h \neq 1, 2, 3\dots$: $C(-h, y) = (-h)!/y!(-y - h)!$.

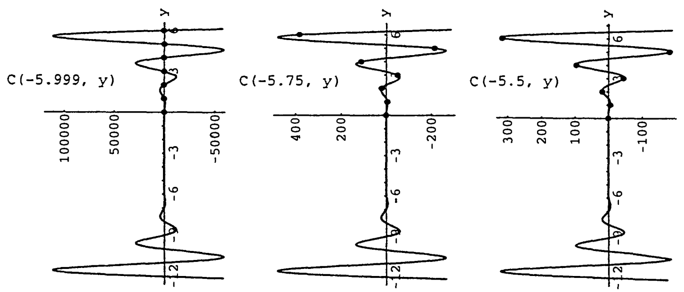
We rewrite the identity above as $\sin \pi y = -\pi/y!(-y - 1)!$; then this section is a generalisation of this, a kind of pseudo-sine. We have decoupled the left- and right-hand zeros of sine and can now move them independently: this pseudo-sine has zeros at $-1, -2, -3, \dots$ and $1 - h, 2 - h, 3 - h, \dots,$ is bounded for $h \leq 1$, unbounded for $h > 1$, and the closer h is to 1, the closer is the analogy to sine; see the sections $C(-\frac{1}{2}, y)$ and $C(-\frac{3}{2}, y)$. The section $C(-1, y)$ is indeed a multiple of $\sin y$, but the multiplier $(-1)!$ is not defined; and the nearby sections $C(-1 \pm \eta, y)$, η small, are large positive or negative multiples of nearly sine functions. When $h \ll 1$, the zero-less gap in the middle increases, and the oscillations of the function decrease very rapidly as $|y|$ increases: so, for example, when $x = n + \frac{1}{2}$, the function has zeros at $-1, -2, -3, \dots$ and $n + \frac{3}{2}, n + \frac{5}{2}, n + \frac{7}{2}, \dots,$ thus making room for Pascal's ridge in the interval $[0, n + \frac{1}{2}]$, and it decreases rapidly to zero outside the interval $[-1, n + \frac{3}{2}]$. On the other hand, when $x = -n - \frac{1}{2}$, it has zeros at $-1, -2, -3, \dots$ and $-n + \frac{1}{2}, -n + \frac{3}{2}, -n + \frac{5}{2}, \dots$ which overlap to give a cluster of zeros at $-n, -n + \frac{1}{2}, -n + 1, -n + \frac{3}{2}, -n + 2, \dots, -1$ and $-\frac{1}{2}$. For $n \gg 1$, this corresponds to the very-close-to-zero behaviour in the octant $-1 > y > x + 1$ away from the lines $x = -1, -2 -3, \dots$. Some sections shows the behavior at other points than $-n - \frac{1}{2}$.

Vertical sections $C(h, y)$: read the graphs sideways!

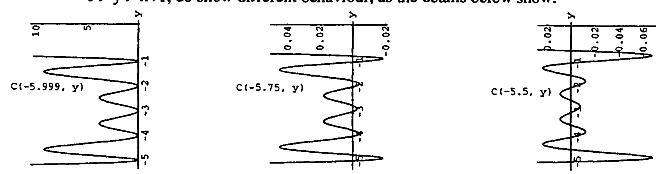
(The non-zero binomial coefficients in $(a+b)^h$, $h \neq -1, -2, \ldots$, are indicated.)



Typical sections for $h << 0$, with different fractional parts.
The sections $C(-n+0, 5\pm\eta)$ do not appear to differ greatly from each other, apart from their vertical scale, and a stretch through h for $y > 0$:

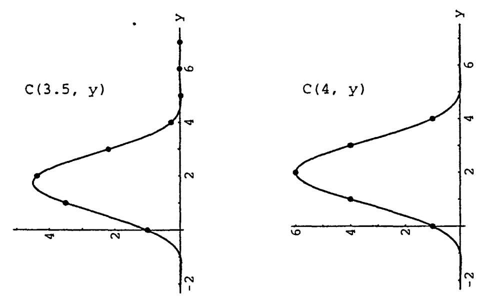


However the central, very close-to-zero sections, which come from $-1 > y > x+1$, do show different behaviour, as the details below show:



and $C(-6.001, y) \approx C(-5{,}999, y)$; their difference is imperceptible at this scale.

Typical sections for $h >> 0$, h non-integral, and h integral



Observe how, at points like $(-6\pm.001, -2.5)$, very close to and on opposite sides of the line $x = -6$ of singularities, the function takes values less than 4. By contrast, at $(-6\pm.001, -11.5)$, it takes values of the order of 10^5.

Horizontal sections.

$y = 0$**:** $C(x, 0) = x!/x! = 1$.

$y = m$**:** $C(x, m) = x!/m!(x - m)! = x(x-1)\dots(x-m+1)/m!$, the binomial coefficient for *all* real x. These polynomials are zero on the integer lattice in the octant $0 \le x < y$; they behave like x^m for large $|x|$; and they are symmetric (n even) or antisymmetric (n odd) about the point $x = \frac{1}{2}(m-1)$.

$y = -m$**:** $C(x, -m) = x!/(-m)!(x+m)! = 0$.

$y = \pm\eta$**:**$C(x, \pm\eta) = x!/\eta!(x \mp \eta)!$, the basic functions underlying all other sections $y = \pm m \pm \eta$. The graph for the cases $\pm\frac{1}{2}$ are given in the collection of horizontal sections; they have zeros at $(-\frac{1}{2}), -\frac{3}{2}, -\frac{5}{2}, \dots$.

$y = m + \eta$**:**

$$C(x, m+\eta) = \frac{x!}{(\eta+m)!(x-\eta-m)!}$$
$$= \frac{(x-\eta)(x-\eta-1)\dots(x-\eta-m+1)}{(\eta+1)(\eta+2)\dots(\eta+m)} C(x,\eta).$$

This introduces extra zeros at $\eta, \eta+1, \dots, \eta+m-1$, thus making the function very close to zero in the interval $[\eta, m+\eta-1]$, in fact in $[\eta-1, m+\eta-1]$, because $C(x,\eta)$ already has a zero at $\eta - 1$. A range of sections for the case $\eta = \frac{1}{2}$ is given, and also a sequence of sections round $y = 3$, where we see a complicated function with infinities is distorted through a simple polynomial into a different complicated function.

$y = -m - \eta$**:**

$$C(x, -m-\eta) = \frac{x!}{(-\eta-m)!(x+\eta+m)!}$$
$$= (-1)^m \frac{\eta(\eta+1)\dots(\eta+m-1)}{(x+\eta+1)(x+\eta+2)\dots(x+\eta+m)} C(x, -\eta).$$

Here the apparent new singularities at $-\eta-1, -\eta-2, \dots, -\eta-m$ in fact cancel the corresponding zeros of $C(x, -\eta)$, generating nearby maxima or minima in these intervals $(-m-1, -m)$.

Diagonal sections. First, sections by lines of positive slope:

$y = x + h$: $C(x, x+h) = x!/(x+h)!(-h)! = C(x, -h)$,

so these correspond to the horizontal sections, described earlier.

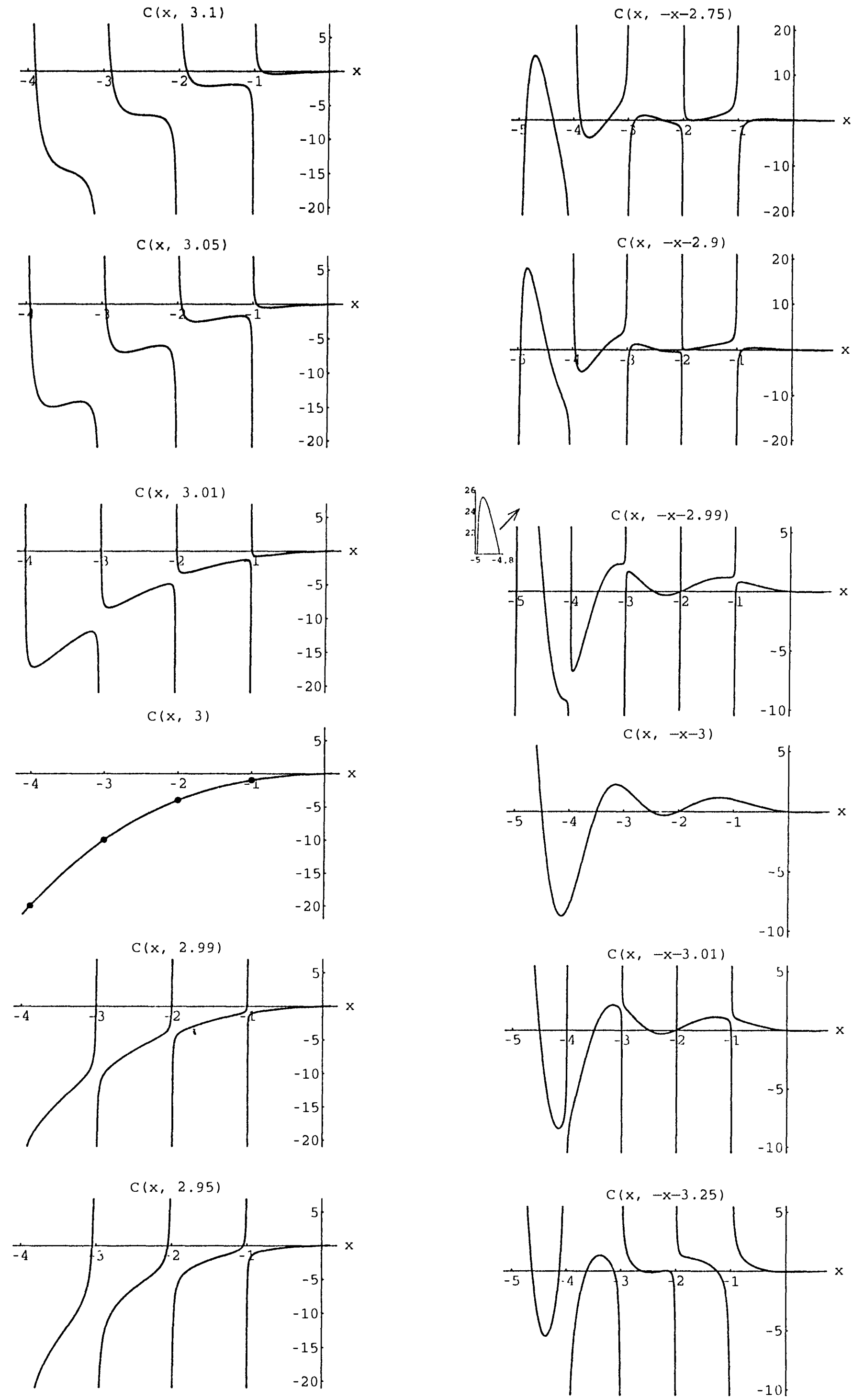
Next, lines of negative slope. Here we can appeal to the duplication formula, $(2x)! = (1/\sqrt{\pi})2^{2x}x!(x-\frac{1}{2})!$, for the simplest case:

$y = -x$**:** $C(x, -x) = x!/(-x)!(2x)! = \sqrt{\pi}/2^{2x}(x-\frac{1}{2})!(-x)!$. This is again a pseudosine (see $C(-h, y)$), here with zeros at $1, 2, 3, \dots$ and $-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots$, multiplied by the negative exponential 2^{-2x} which makes it increase rapidly for

Moving through an integer point

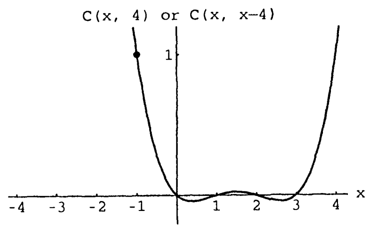
Down through $C(x, 3)$, or up through $C(x, x–3)$, in detail:

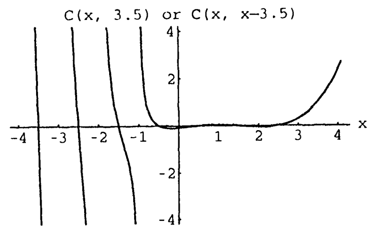
Down through $C(x, –x–3$ in detail:

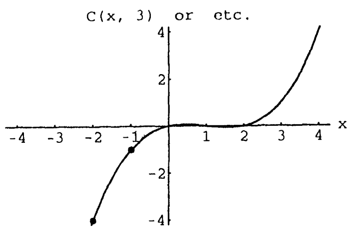


Horizontal sections $C(x, h)$, or diagonal sections $C(x, x-h)$

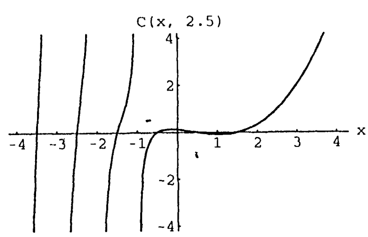
(The non-zero coefficients in $(a+b)^{-n}$ are indicated.)

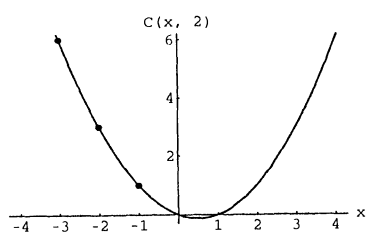


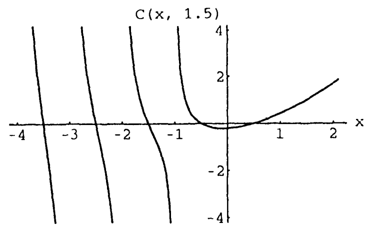




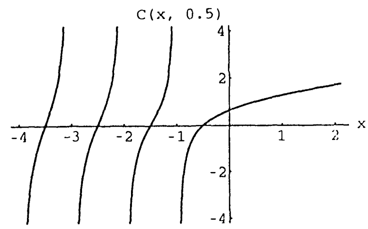
(A further sequence of sections around $C(x, 3)$ is given below.)



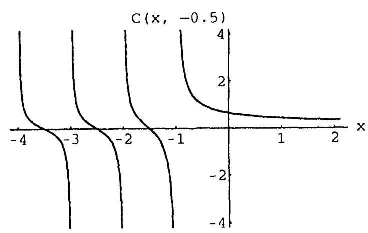


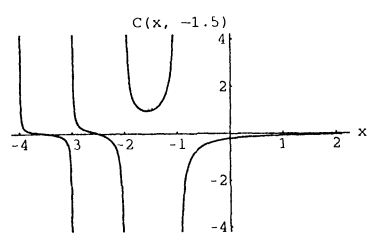


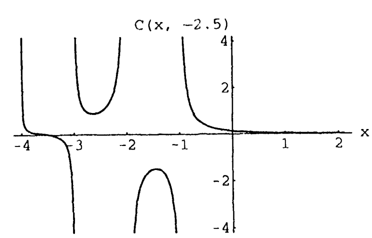
$C(x, 1) = x$



$C(x, 0) = 1$, and $C(x, -n) = 0$ for all n

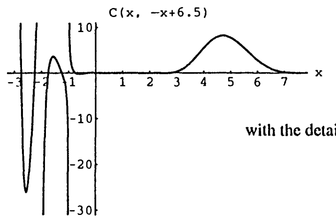




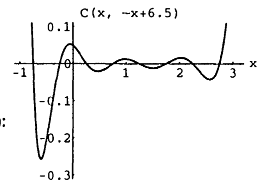


Diagonal sections of negative slope $C(x, -x+h)$

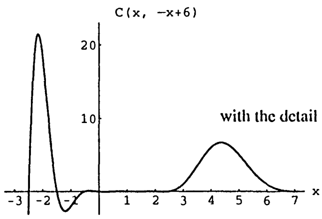
Typical sections for $h >> 0$, h non-integral:



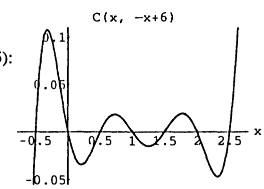
with the detail of (–1, 3):



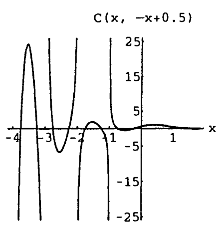
... and h integral:

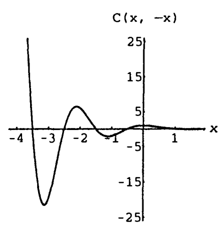


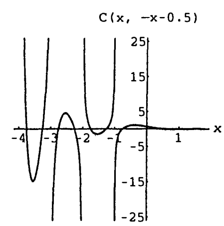
with the detail of (–0.5, 2.5):



... near the origin:

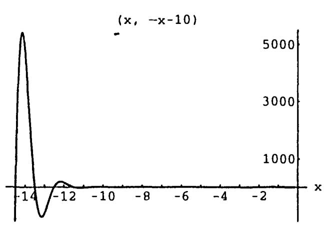






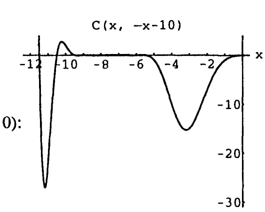
... for h around –3 (see the separate sequence),

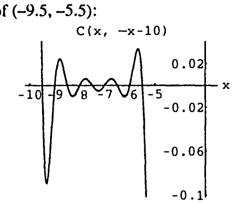
... and for $h << 0$, h integral:



and further detail of (–9.5, –5.5):

with the detail of (–11.5, 0):





$x < 0$, and decrease rapidly for $x > 0$. The sections $y = -x \pm n$ can then be expressed in terms of this.

$y = -x \pm n \pm \eta$: Although I can see no similar simplifications for these general cases, there is no difficulty in applying the now familiar manipulations to identify the singularities and zeros of the functions. A case of particular interest is:

$y = -x - n - \eta$: $C(x, -x - n - \eta) = x!/(-x - n - \eta)!(2x + n + \eta)!$.

This has singularities at $x = -1, -2, -3, \ldots$ and zeros when $-x - n - \eta = -1, -2, -3, \ldots$, i.e. $x = -n + 1 - \eta, -n + 2 - \eta, -n + 3 - \eta, \ldots$, and when $2x + n + \eta = -1, -2, -3, \ldots$, i.e. $x = \frac{1}{2}(-n - 1 - \eta)$, $\frac{1}{2}(-n - 2 - \eta)$, $\frac{1}{2}(-n - 3 - h), \ldots$. Consider first the case of $y = -x - 2\frac{3}{4}$: zeros at $-1\frac{3}{4}, -\frac{3}{4}, \frac{1}{4}, 1\frac{1}{4}, \ldots$ and at $-1\frac{7}{8}, -2\frac{3}{8}, -2\frac{7}{8}, -3\frac{3}{8}, \ldots$. Hence in every interval $[-2, -1], [-3, -2], [-4, -3], \ldots$ between the singularities, there are two zeros, an even number, so the function must tend to infinity in the same direction at each end-point of these intervals and have an odd total number of maxima and minima between. This behavior is typical of each section $y = -x - h, h < 3$ and non-integral. Now consider $y = -x - 3\frac{1}{4}$: zeros at $-2\frac{1}{4}, -1\frac{1}{4}, \frac{3}{4}, 1\frac{3}{4}, \ldots$ and at $-2\frac{1}{8}, -2\frac{5}{8}, -3\frac{1}{8}, -3\frac{5}{8}, \ldots$. Here $[-2, -1]$ contains one zero, $[-3, -2]$ three zeros, then the following intervals contain an even number of zeros. Hence the first two of these intervals will now contain an even total number of maxima and minima, and the function will tend to infinity in opposite directions at their end-points. This kind of transition is illustrated here in a remarkable sequence of sections around $y = -x - 3$, and is repeated at each succeeding section $y = -x - n, n > 3$.

4. A HISTORICAL NOTE. The configuration of values of this function on the positive integer lattice in the octant $0 \leq y \leq x$ may now be called Pascal's triangle in the west, but it is well known that it is attested long before Pascal's *Traité du Triangle* was compiled and distributed posthumously in 1665 from sheets that had been printed in 1654. For further details and references concerning most of the names in the summary below, see the *Dictionary of Scientific Biography* (= *DSB*; references will be cited by acronym), whose transliterations and spellings I have adopted, and Edwards, *PAT*, which also contains some illustrations from original sources.

Knowledge of the expansions of $(a + b)^2$ and $(a + b)^3$ are very old indeed, but the first surviving reference to a more general result seems to be in the Arabic mathematician al-Karajī, sometime around 1100. He described how to evaluate the terms in the binomial expansions up to the fourth degree, with proof, then the fifth degree, with a statement only, and then 'and so on', all by a process that can be generalised but which does not quite have the simplicity and transparency of the arithmetical triangle. Then, in reporting al-Karajī's work, al-Samaw'al (†1175) described the generating rule $\binom{n}{m} = \binom{n-1}{m-1} + \binom{n}{m-1}$, together with a table up to the 12th degree; see Rashed, IMKS. The procedure was perhaps known by al-Khayyāmī, aka Omar Khayyam, (1039–1123); it is found in al-Ṭūsī (13th century); and in al-Kāshī (15th century), where, as often elsewhere, it is used in the calculation not of powers, but of roots; see Berggren, *EMMI*, 53–63 and Rashed, ER.

Binomial coefficients are also explicitly attested early in China, again in connection with finding roots. There are surviving illustrations of it in Chu Shih-chieu's *Precious Mirror of the Four Elements* (1303), and in Yang Hui's *Detailed Analysis of*

the Mathematical Rules in the 'Nine Chapters' (1261); Yang says that Chia Hsien had given the rule about 1050; and it was probably also described in an earlier now-lost book by Liu Hui. Reproductions of the surviving illustrations are given in Needham, *SCC* 133ff and Boyer, *HM* 228, and further details in Lǐ & Dù, *CM*.

Needham writes (*SCC* 137, note *a*): "A claim for a great antiquity of the Pascal triangle in India has been made by [A.N.] Singh. But though a similar numerical pattern (the *Meru-prastāra*) was derived from a passage in Piṅgala's *Chandaḥ-sūtra*, VIII 23 (c. 200 BC) by the 10th century commentator Halāyudha, it concerns prosodic combinations only and has nothing to do with binomial coefficients"; but this surely is too restrictive a view. It is true that the Indian interest is mainly in the patterns that can arise in Sanskrit prosody (though not exclusively; for example Brahmagupta (AD 629) did consider $(a + b)^3$), but the western studies of figured numbers, binomial coefficients, and combinations are, ultimately, a search for the same patterns and yields the same collections of numbers. (For another illustration involving the Fibonacci numbers, see P. Singh, SCFN.) Here, for example, is Halāyudha's description of how to generate an enumeration of these syllabic combinations: "Draw a square. Beginning at half the square, draw two other similar squares below it; below these two, three other squares, and so on. The marking should be started by putting 1 in the first square. Put 1 in each of the two squares of the second line. In the third line put 1 in the two squares at the ends and, in the middle square, the sum of the digits in the two squares lying above it. In the fourth line put 1 in the two squares at the ends. In the middle ones put the sum of the digits in the two squares above each. Proceed in this way. Of these lines, the second gives the combinations with one syllable, the third the combinations with two syllables, etc."—or, we would say, they also give the binomial coefficients. For more details, see Edwards, *PAT* 27ff, from which this quotation has been taken.

In the west, the tableau of numbers appeared on the title page of the book *Rechnung* (1527) by the German algebrist Apian (for a reproduction, see Boyer, *HM* 328). Then there is a table up to 18th order in another German book, Stifel's *Arithmetica Integra* (1544), then in his *Deutsche Arithmetica* (1545) and his new edition of Rudolff's *Coss* (1545, originally published in 1525) and, according to the *DSB* article by Kurt Vogel, Stifel reports that "he discovered these coefficients with great difficulty, having found no one to teach them to him nor any written accounts of them." Thereafter, it seems generally well-known: it or closely related material is found in Scheubel (1545), Peletier (1549), Tartaglia (1556; the arithmetical triangle is sometimes known in Italy as Tartaglia's triangle), Cardan (1570; see Boyer CPT, the inspiration for this list), Stevin (1585), Faulhaber (1615), Girard (1629), Oughtred (1631), Briggs (1633), Mersenne (1636), and Fermat (1636), to bring the list up to Pascal and stop there, but with no claim to completeness. It was then Montmort (1708) who referred to the "Table de M. Pascal pour les combinaisons" and De Moivre (1730) to the "Triangulum Arithmeticum PASCALIANUM," and the name has stuck, a fine example of what Merton, *OTSOG* 218, describes as how "in the transmission of ideas each succeeding repetition *tends to erase* all but one antecedent version, thus producing what may be described as the anatopic or palimpsest syndrome."

However we, here, are not really interested in the arithmetical triangle in itself, but in its interpolations and extrapolations, and it is to these we now turn. But before that, a brief digression. The origins and first uses of mathematical induction have been vigorously debated (for my own contribution, see Fowler, GMI), but the first explicit statement and use of the principle itself was always thought to be in

Pascal's pamphlet; it has now been proposed in Rashed, IMKS (this article also gives a survey of the literature on the topic) that al-Samaw'al should also have some credit for this. And these are two of the texts that have framed our discussion so far!

Sometime around or after 1600 (the dating is problematic), Harriot wrote a substantial treatise *Magisteria Numerorum Triangularium* or *De Numeris Triangularibus et inde De progressionibus Arithmeticis Magisteria magna* on these coefficients, which he used for interpolation in general. In this, he considered $\binom{h}{m}$ with negative and fractional values of h. This treatise has never been published, but there is a brief reference to it and some examples from it in Lohne, ETH 293-4; for example, a table of $\binom{n}{m}$ with $-5 \leq n \leq 4, 0 \leq m \leq 5$. Also see Knuth, RHBN 241, who reports that "[Harriot] computed the polynomials $\binom{n}{k}$ for $k \leq 7$, writing $nnnn - 6nnn + 11nn - 6n/24$ for $\binom{n}{4}$."

Even before Pascal's pamphlet was published, Wallis, in his *Arithmetica Infinitorum* (1656), was, in effect, interpolating further the integer points to find expressions for some half-integer points. In his search for some expression for π or, more accurately, for the ratio of the unit square to the quarter circle of unit radius, $\frac{4}{\pi}$, he guessed and tabulated expressions for

$$W(p,q) = \frac{1}{\int_0^1 (1 - x^{1/p})^q \, dx} = \frac{(p+q)!}{p!q!} = C(p+q,q) = \frac{1}{pqB(p,q)}$$

for non-negative integral values of p and q. (This is a symmetrical version of our binomial coefficient function, symmetric around the line $p = q$, and almost the reciprocal of the classical beta function.) The value he wanted was $W(\frac{1}{2},\frac{1}{2}) = C(1,\frac{1}{2})$, but, in deriving this as an infinite product,

$$W(\tfrac{1}{2},\tfrac{1}{2}) = \tfrac{4}{\pi} = \tfrac{3}{2} \times \tfrac{3}{4} \times \tfrac{5}{4} \times \tfrac{5}{6} \times \tfrac{7}{6} \times \tfrac{7}{8} \times \dots,$$

he also found and tabulated expressions for all of the other half-integer points in the positive octant, indeed for $-\frac{1}{2} \leq y \leq x + \frac{1}{2} \leq 5, -1 \leq x - y \leq 5$, with the value in its corner, $W(-\frac{1}{2},-\frac{1}{2}) = C(-1,-\frac{1}{2})$, that he gave as ∞, the symbol introduced by him and the only previously published hint I know of anywhere of the line of singularities of C on $x = -1$; see Plate 1. (He also says, in passing, in his Proposition 184, that he could easily continue his procedure to find the interpolations to thirds, quarters, or more points in each interval, and thus he could have plotted a dense set of points on our surface!) Wallis also reported Brouncker's continued fraction expression for the same quantity:

$$\frac{4}{\pi} = 1 + \cfrac{1}{2 + \cfrac{9}{2 + \cfrac{25}{2 + \cfrac{49}{2 + \ddots}}}}$$

and its derivation. Throughout his explorations, Wallis referred to his table as being made up of figured numbers, and later, in his *Treatise of Algebra* (1885), Chapter 85, he explicitly denied any connection with the binomial theorem: "For, [the expansion] for Cubicks consisting of 4 Members; that for Quadraticks, of 3; that for Laterals, of 2; and that for Equals, of a Single Unite; and suitably for other Powers: That for the Quadratick Root of Laterals [i.e. $(a + b)^{1/2}$], should, by this

Prop. 189. *Arithmetica Infinitorum.* 169

quem determinate numerum designet (et propterea nihil inde certi de reliquis quantitatibus concludi possit,) potest tamen quasi virtualiter cujusvis numeri vices subire. Nam quicunq; numerus per ∞ dividatur, quotientem dabit 0. et contra Puta $\infty)1(0.0)1(\infty.\infty)2(0.0)2(\infty.\infty)3(0.0)3(\infty$. Et sic de quovis alio: Et propterea (cum *Divisor* in *Quotientem* ductus restituere debeat numerum *Dividendum*) esset $\infty \times 0 = 1$, vel $\infty \times 0 = 2$, vel $\infty \times 0 = 3$; Et sic de quovis alio numero.

PROP. CLXXXIX. *Theorema.*

Hinc sequitur, quod Si ex Tabellæ prop. 184. locis vacuis unus quilibet numero noto suppleatur, erunt & reliqui omnes cogniti.

Verbi gratiâ; si numerus hâc notâ $\square$ designatus supponatur cognitus, reliqui omnes etiam cognoscentur; qui nempe eam habent ad illum rationem quæ hic subtus indigitatur.

∞	1	$\frac{1}{2}\square$	$\frac{1}{2}$	$\frac{1}{3}\square$	$\frac{3}{8}$	$\frac{4}{15}\square$	$\frac{15}{48}$	$\frac{8}{35}\square$	$\frac{105}{384}$	A
1	1	1	1	1	1	1	1	1	1	1
$\frac{1}{2}\square$	1	$\square$	$\frac{3}{2}$	$\frac{4}{3}\square$	$\frac{15}{8}$	$\frac{8}{5}\square$	$\frac{105}{48}$	$\frac{64}{35}\square$	$\frac{945}{384}$	$A \times \frac{2l-1}{1}$
$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	$\frac{7}{2}$	4	$\frac{9}{2}$	5	$l = \frac{2l \mp 0}{2}$
$\frac{1}{3}\square$	1	$\frac{4}{3}\square$	$\frac{5}{2}$	$\frac{8}{3}\square$	$\frac{35}{8}$	$\frac{64}{15}\square$	$\frac{315}{48}$	$\frac{128}{21}\square$	$\frac{3465}{384}$	$A \times \frac{4l^2-1}{3}$
$\frac{3}{8}$	1	$\frac{15}{8}$	3	$\frac{35}{8}$	6	$\frac{63}{8}$	10	$\frac{99}{8}$	15	$\frac{l^2+l}{2} = \frac{4l^2+4l}{8}$
$\frac{4}{15}\square$	1	$\frac{8}{5}\square$	$\frac{7}{2}$	$\frac{64}{15}\square$	$\frac{63}{8}$	$\frac{128}{15}\square$	$\frac{693}{48}$	$\frac{512}{35}\square$	$\frac{9009}{384}$	$A \times \frac{8l^3+12l^2-2l-3}{15}$
$\frac{15}{48}$	1	$\frac{105}{48}$	4	$\frac{315}{48}$	10	$\frac{693}{48}$	20	$\frac{1287}{48}$	35	$\frac{l^3+3l^2+2l}{6} = \frac{8l^3+24l^2+16l}{48}$
$\frac{8}{35}\square$	1	$\frac{64}{35}\square$	$\frac{9}{2}$	$\frac{128}{21}\square$	$\frac{99}{8}$	$\frac{512}{35}\square$	$\frac{1287}{48}$	$\frac{1024}{35}\square$	$\frac{19305}{384}$	$A \times \frac{16l^4+64l^3+56l^2-16l-15}{105}$
$\frac{105}{384}$	1	$\frac{945}{384}$	5	$\frac{3465}{384}$	15	$\frac{9009}{384}$	35	$\frac{19305}{384}$	70	$\frac{l^4+6l^3+11l^2+6l}{24} = \frac{16l^4+96l^3+176l^2+96l}{384}$

A a a Totus

Plate 1. The complete interpolated table in Proposition 189 of J. Wallis, *Arithmetica Infinitorum* (1656). Reproduced by permission of the Bodleian Library, Oxford, from J. Wallis, *Operum Mathematicorum pars altera*, Savile K 20, p. 169, Wallis' own copy of the first edition.

analogy, consist of more Members than one, but fewer than Two." For more details, see Whiteside, PMT 236ff. and Edwards, *PAT* 87–95.

The solution of this difficulty, as Harriot already knew and Newton was soon to exploit, is that the arithmetical triangle should float in a sea of zeros, so that most coefficients in these expansions are zero. As Newton was to explain later (see Newton, *MP*; the following quotation comes from p.8, n. 22 and the calculations on Plate 2 are edited on pp. 122–6): "In the winter between the years 1664 & 1665 upon reading Dr Wallis's Arithmetica Infinitorum & trying to interpole his progressions for squaring the circle, I found out first an infinite series for squaring the circle & then another infinite series for squaring the hyperbola & soon after." This was Newton's discovery of the general binomial theorem and, on the way to it, he wrote down tables of binomial coefficients $\binom{n}{m}, -4 \le n \le 7, 0 \le m \le 7$, and first did a polynomial interpolation of them (or evaluated Wallis' polynomial expressions; see Plate 1) at the points $(n + \frac{1}{2}, m)$; and since $\binom{n+1/2}{m}$ *is* a polynomial of appropriate degree, this gives the precise value at these points. He then went on to interpolate at the points $(n + \frac{1}{3}, m)$ and $(n + \frac{2}{3}, m)$, where a similar thing happens; indeed, for any rational (or even real) r, the root binomial coefficient $\binom{r}{1}$ is r, so this procedure will give the precise value for $\binom{n+r}{m}$. (It should be emphasised that our simple and universally known expressions $\binom{n}{m} = n(n-1)\dots(n-m+1)/m(m-1)\dots1 = n!/m!(n-m)!$ were only just emerging at this time though, as we have seen, particular cases of these formulae are found in Harriot and Wallis.)

The factorial or gamma function (even today, some mathematicians seem to feel a frisson on thinking of the factorial as a function of a real or complex variable; see Fowler, SAFF) was explored by Euler, who called it an 'inexpressible function', a function whose formula is not immediately obvious; another example he considered was the series $1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}$ interpolated to real variables. Euler's investigation of the factorial function started from the infinite product

$$n! = \left[\left(\frac{2}{1}\right)^n \frac{1}{n+1}\right]\left[\left(\frac{3}{2}\right)^n \frac{2}{n+2}\right]\left[\left(\frac{4}{3}\right)^n \frac{3}{n+3}\right]\dots$$
$$= \lim_{m\to\infty} \frac{m!(m+1)^n}{(n+1)(n+2)\dots(n+m)}$$

which, he noticed after further manipulation, yielded Wallis' product for $n = \frac{1}{2}$; and he manipulated it yet further into

$$n! = \int_0^1 (-\log x)^n \, dx;$$

and all of these formulae make sense for any positive real n. (The translation of the Latin title of his first article on the subject was: 'On transcendental progressions whose general term cannot be expressed algebraically'; we can interpolate the sequence of triangular numbers easily with the function $t(x) = \frac{1}{2}n(n+1)$, but we cannot find such an algebraic interpolation for the factorials, for example because $\frac{1}{2}! = \sqrt{\pi}/2$.) Legendre introduced the name 'gamma function', $\Gamma(x) = (x-1)!$; Gauss referred to $\pi(x) = x!$; Weierstrauss wrote $Fc(x) = (x!)^{-1}$, which he called the *factorielle*, an important entire function with half of the zeros of the sine, as we have seen; and other names have been used. There is a nice leisurely account of this and associated topics in Davis, LEI, to which I commend the reader, and the rich story of notation and nomenclature can be filled in from Cajori, *HMN*.

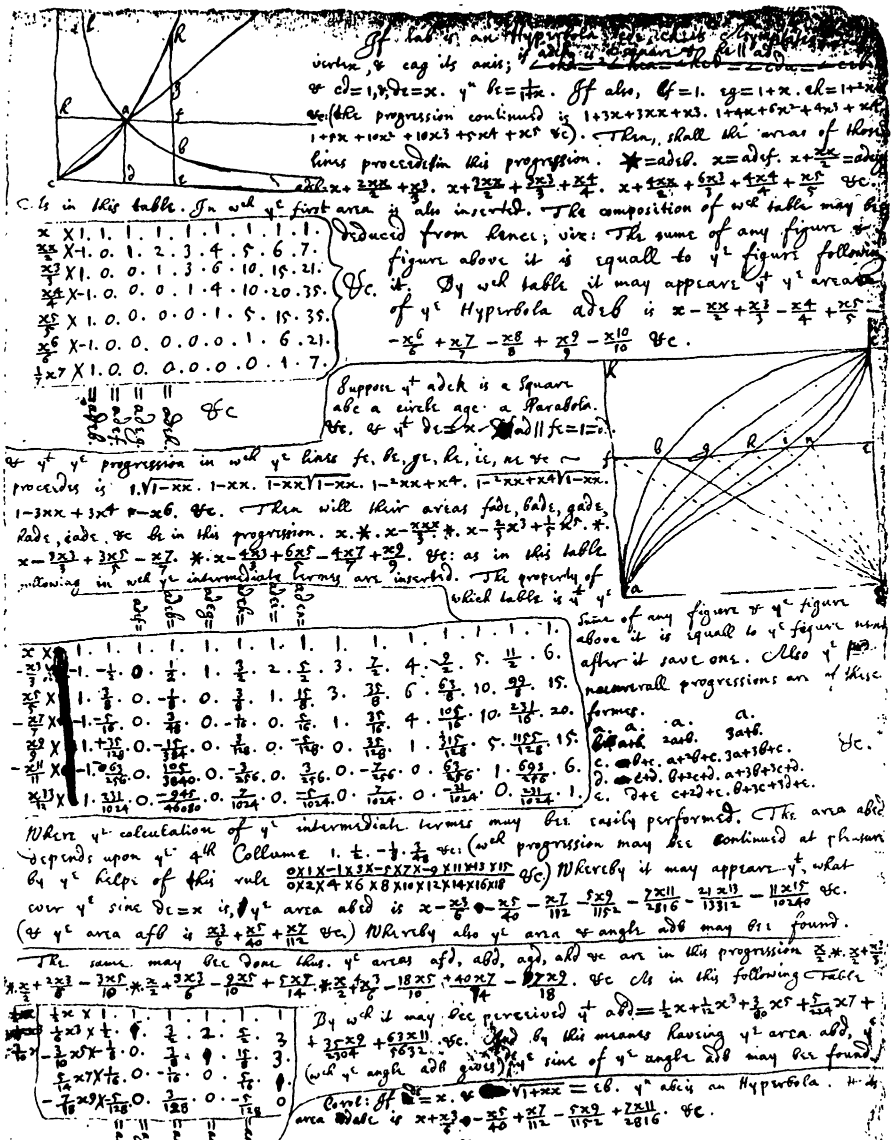


Plate 2. A leaf of Newton's *Annotations out of Dr. Wallis his Arithmetica infinitorum.* Reproduced by permission of the Syndics of Cambridge University Library from Cambridge University Library Add. 3958.3:72r.

I finish with a puzzled observation: I know of no evidence that the graph of C has ever been plotted before, and have some evidence that few think of the possibility of $\binom{x}{y}$ being a function of two real variables. For example, very few people have been able to recognise the surface, even with the inducement of a substantial prize, though several have come back to me after they have been told and said that they 'really' knew it all along; some have even sent photocopies of notes showing that they had previously drawn sections of it, though never any involving discontinuities. (I have always believed that mathematicians are revisionist, but have never seen such a striking example as this.) Further, the Binomial[x, y] function in the usually very comprehensive *Mathematica*® requires $(x - y)$ to be an integer, thus leading to the polynomial expressions for the horizontal sections $C(x, \pm m)$ above, an eminently practical approach that neatly finesses the discontinuities of the function, though of course its very closely associated Beta[x, y] is a pukka function of two real variables with its own complicated set of discontinuities. The description in Graham, Knuth, and Patashnik, *CM* 211, that "A binomial coefficient can be written $\binom{z}{w} = \lim_{\zeta \to z} \lim_{\omega \to w} \zeta!/\omega!(\zeta - \omega)!$ when z and w are any complex numbers whatever," is a succinct way of expressing the horizontal directional limits at $(-m, \pm n)$, but it fails to make explicit that other directional limits at these points are different and that the limits at other points $(-m, h)$ will be the point at infinity on the Riemann sphere, and it gives no idea of the complexity of the Avenues of Manhattan. (This book contains a stimulating and comprehensive account of things binomial.) Also, I know of no earlier explanation of the phenomenon $\binom{n}{m} = \binom{n}{n-m}$ but $\binom{-n}{m} \neq \binom{-n}{-n-m}$. However, I may well be wrong: *Maple*® does define the binomial function for all real $x \neq -1, -2, -3, \ldots$ and y; and Anthony Edwards has referred me to statistical investigations by Karl Pearson of the 'point binomial distribution' which use sections of the surface by $x = h$, $h > 0$ (see Pearson, CMTE).

The idea of this article came to me one day as, on my way into work, I was wondering about Knuth's elementary and sensible practice of writing the binomial theorem as $(a + b)^x = \sum_{m \in Z}\binom{x}{m}a^m b^{x-m}$, where either x is a non-negative integer or $|a/b| < 1$; so what did the graph of $\binom{x_0}{y}$ look like for different values of x_0? Or $\binom{x}{y_0}$ for different values of y_0? Instead of working these out analytically, I asked Mark Muldoon to get his computer to draw out the whole surface, with the idea at that stage of submitting a one-page article—this surface and four lines of explanation—on this to the *Monthly*. But the surface turned out to be so much more complicated than I had anticipated that I started exploring it, trying to get colleagues to identify it (with almost no success, even with lots of clues and help) and interest them in its peculiarities. Tom Whiteside joined in with energy, enthusiasm, and erudition, rekindling our joint interest in Wallis' explorations and our exasperation with his expository style; and Anthony Edwards filled in some details. I then tried to get colleagues, friends, or family to do a lot of plotting of associated graphs, without any success until Jeff Smith told me that, though he wouldn't do them for me himself, he would initiate me in Unix and Windows, so that I could do them for myself—the best bit of instruction I have had in many a year. Fortunately Mark Muldoon soon went backpacking in the Sierras and lent me his Macintosh, a much more user-friendly set-up. Stan Wagon gave me some crucial advice on improving my amateur graphics, later followed up by further help from Tom Wickham-Jones, John Rawnsley, and Igor Rivin; I can send the code of

the *Mathematica* plots to anyone who is interested in exploring and improving them further.

REFERENCES

J. L. Berggren, *EMMI = Episodes in the Mathematics of Medieval Islam*, Springer, 1986.

C. B. Boyer, CPT = Cardan and the Pascal triangle, *American Mathematical Monthly* 57 (1950) 387–91.

C. B. Boyer, *HM = A History of Mathematics*, Wiley, 1968.

F. Cajori, *HMN = A History of Mathematical Notations*, vol 2: Notations mainly in higher mathematics, Open Court, 1929.

P. J. Davis, LEI = Leonhard Euler's Integral: A historical profile of the gamma function, *American Mathematical Monthly* 66 (1959) 849–869.

DSB = Dictionary of Scientific Biography, ed. C. C. Gillispie, Scribner's, 1970–8. The mathematical entries have been extracted in the *Biographical Dictionary of Mathematicians*, Scribner's, 1991.

A. W. F. Edwards, *PAT = Pascal's Arithmetical Triangle*, London: Charles Griffin & New York: Oxford University Press, 1987.

D. H. Fowler, GMI = Could the Greeks have used mathematical induction? Did they use it?, *Physis* 31 (1994) 253–265, with a reply by S. Unguru, Fowling after induction, *ibid.* 267–272.

D. H. Fowler, SAFF = A simple approach to the factorial function, *Mathematical Gazette*, to appear.

R. L. Graham, D. E. Knuth, & O. Patashnik, *CM = Concrete Mathematics*, Addison-Wesley, 1989.

D. E. Knuth, RHBN = Review of A. Glaser, *History of Binary and other Non-Decimal Numeration*, *Historia Mathematica* 10 (1983) 236–243.

Lǐ Yǎn & Dù Shıran, *CM = Chinese Mathematics, A Concise History*, tr. J. N. Crossley & A. W.-C. Lun, Oxford University Press, 1987.

J. A. Lohne, ETH = Essays on Thomas Harriot, *Archive for History of Exact Sciences* 20 (1979) 189–312.

R. K. Merton, *OTSOG = On the Shoulders of Giants, A Shandean Postcript*, The Free Press 1965, and often reprinted, most recently in The Post-Italianate Edition with Foreword by U. Eco, Afterword by D. Donoghue, and Preface and Postface by the author, University of Chicago Press, 1993.

J. Needham, *SCC = Science and Civilisation in China*, vol. 3: Mathematics, etc., Cambridge University Press, 1959.

I. Newton, *MP = The Mathematical Papers of Isaac Newton*, vol. 1, ed. D. T. Whiteside, Cambridge University Press, 1967.

K. Pearson, CMTE = Contributions to the mathematical theory of evolution II: Skew variation in homogeneous material, *Philosophical Transactions of the Royal Society* 186A (1895) 343–414.

R. Rashed, IM = L'induction mathématique: al-Karaǰ, as-Samaw'al, *Archive for History of Exact Sciences* 9 (1972) 1–21, repr. pp. 71–91 of R. Rashed, *Entre Arithmétique et Algèbre, Recherches sur l'Histoire des Mathématiques Arabes*, Les Belles Lettres, 1984.

R. Rashed, ER = L'extraction de la racine $n^{\text{ième}}$ et l'invention des fractions décimales (XI^{e}—$\mathrm{XII}^{\mathrm{e}}$ siècles), *Archive for History of Exact Sciences* 18 (1978) 191–243, repr. pp. 93–145 of R. Rashed, *Entre Arithmétique et Algèbre, Recherches sur l'Histoire des Mathématiques Arabes*, Les Belles Lettres, 1984.

P. Singh, SCFN = The so-called Fibonacci numbers in ancient and medieval India, *Historia Mathematica* 12 (1985) 229–244.

J. Wallis, *AI = Arithmetica Infinitorum, sive Nova Methodus Inquirendi in Curvilineorum Quadraturam, aliaque difficiliora Matheseos Problemata*, Oxford, 1656 = *Operum mathematicorum pars altera*, London, 1656, and repr. in *Opera Mathematica* 1 (1695).

D. T. Whiteside, PMT = Patterns of mathematical thought in the later seventeenth century, *Archive for History of Exact Science* 1 (1961) 179–388.

Mathematics Institute
University of Warwick
Coventry CV4 7AL
ENGLAND
dhf@maths.warwick.ac.uk

Classical Mathematics.
Baroque Mathematics.
Romantic Mathematics?
Mathematics *Jazz!*
Also Atonal, New Age, Minimalist, and Punk Mathematics.

Jim Henle[1]

Mathematicians love music. Nearly everyone likes music, of course, but mathematicians are typically extreme. Many are devotees. More than a few are amateur musicians of some skill and much energy. Some seem to love music more than their own discipline. Why?

The two are poles apart. Compare: science on the one hand, art on the other; reason versus passion, utility versus felicity. Then too, the nerd versus the rock star. Is it a case of opposites attracting?

Not at all. Certainly the feelings are not reciprocated. Musicians are not famous for their appreciation of mathematics. The few exceptions (academic composers who base works on esoteric formulae) only reinforce our instinct that the music we revere cannot be reduced to algorithms. A crude survey affirms these impressions. At Smith College (no distribution requirements) between 1987 and 1992, mathematics majors were 11% more likely to have taken a music course than the ordinary undergraduate. At the same time, music majors at Smith were 42% *less* likely to have taken a mathematics course than the ordinary undergraduate.

The affinity, however, is thousands of years old. From the Pythogoreans to Aristoxenus to Boethius, philosophers have imagined music as a branch of mathematics. In theory, there is much that is mathematical in music: the physics of sound, the arithmetic of rhythm, the algebra of scales. Is this the answer? Do mathematicians see music as an intellectual cousin?

I think not. Painting, sculpture and architecture all have their mathematical components, yet there is no mystical (or statistical) connection. Smith mathematics majors are 10% *less* likely to have taken a literature course and 15% *less* likely to have taken an art course than the ordinary undergraduate. I would argue, in fact, that cause and effect have been confused here. The existence of countless mathematical analyses of music is merely evidence that mathematicians have been around, picking at the corpus of music in an attempt to understand its appeal.

What then, is the explanation?

Thesis: In brief, it's not that music is *mathematical.* It's that mathematics is *musical.*

I contend that there is something profoundly similar about mathematics and music. I apprehend this by the way the disciplines respond to the intellectual

[1] I would like to acknowledge my colleagues at Smith and elsewhere who were kind enough to read early versions of this paper and call attention to my excesses: Gian-Luigi Bellini, Dan Isaacson, Marjorie Senechal, Ruth Solie, and Tom Tymoczko. Of course, none of these can be held responsible for any such excesses that remain.

currents in society. Over centuries, the patterns of their growth are remarkably alike. Specifically, I will argue:

1. Mathematics has many of the characteristics of an art.
2. Viewed as an art, it is possible to identify artistic periods in mathematics: Renaissance, Baroque, Classical, and Romantic.
3. These periods coincide nicely and share many characteristics with the corresponding musical epochs, but *are significantly different* from those of painting and literature.

The structure of this paper follows the outline above. In the first section we discuss briefly the aesthetic nature of mathematics. In the second, we locate the mathematics that corresponds to the major artistic periods. In the third section, we date the periods.

As the argument unfolds, the reader may note some leaps and stumbles. We will address them in the fourth section. The concluding sections contain some further speculations and the conclusion.

I emphasize **this is not a mathematical study of music**. Such studies are a staple of mathematicians, musicians and philosophers. While often fascinating, they are not relevant. They do not explain the real affinity between mathematics and music. That affinity is more emotional than procedural, more spiritual than intellectual. It is the cultural context that matters, not abstract principles.

Finally, let me add that I do not deny the validity of analogies between mathematics and poetry, nor do I wish to ignore those artists such as Helaman Ferguson now realizing mathematical ideas in stone and on canvas. I will, however, insist that in a very real but elusive sense mathematics, as a cultural enterprise, is closer to music than to literature, painting, or sculpture.

1. MATHEMATICS AS ART. Mathematicians today readily recognize in their discipline the attributes of art. The usual view of mathematics as a science, as rooted in the real world and as a purveyor of absolute truth has much to recommend it, but misses important features and is arguably false.

First of all, the idea of "absolute truth" has been severely damaged in the last two centuries by the discovery of alternative geometries, alternative number systems, and even alternative truth. While many mathematicians still regard their work as having universal validity, their "universe" has been narrowed and "validity" redefined.

Second, mathematicians are creators. While some mathematical worlds are forced into being by the need to understand some real phenomenon, others are created in the same spirit as poems, symphonies, and temples. They are meant to be admired and enjoyed.

Third, mathematicians exercise considerable taste in their research. Beyond truth, we seek beauty, harmony, and elegance. We have aesthetic standards for our proofs. We have standards for the fields we investigate and the theorems we publish.

Not surprisingly, taste differs among mathematicians. Mathematical aesthetics is quite complex, encompassing a greater emotional range than the public commonly imagines. While one mathematical structure is admired for its symmetry, another is prized for its singularity. Mathematicians are attracted not merely to the beautiful, but also to the grand, the picturesque, and the gothic.

This is not to say that mathematics *is* art. What mathematics *is* is difficult to say. It shares traits with the arts, but also with the sciences, with philosophy, with

theology, with sport, with gastronomy, and so on. I argue here only that mathematics can be viewed as an art in a meaningful way, and that such a perspective will help us to understand mathematics—and understand art as well.

2. MATHEMATICAL PERIODS. A word on our procedure. We are going to apply the concepts of "Renaissance," "Baroque," etc. to mathematics. The meaning of these terms have evolved over centuries. We will not use the most authoritative versions, but popular ones, in fact, we choose as our references standard texts in music, art, literature, and mathematics (respectively, Grout: *History of Music*, Gardner: *Art Through the Ages*, Benet: *The Reader's Encyclopedia*, and Kline: *Mathematical Thought from Ancient to Modern*). This is appropriate. Any categorization of artistic periods is arbitrary, but our sources are contemporary and are products of the same cultural context, and context is what this study is all about.

Renaissance Mathematics. There was an identifiable Renaissance in mathematics. It was made possible by the recovery and revival of Greek mathematics, fueled by the rise of commerce, and marked by increasing secularization. It is characterized by work that goes beyond annotation and summarization of the extant classics.

Summa de Arithmetica, Geometria, Proportioni et Proportionalita, Pacioi, 1494, is what we would call typical pre-Renaissance. It organized all that was currently known and issued gloomy predictions for future progress. *Ars Magna*, Cardano, 1545, on the other hand, broke new ground in a subject which classical authors had generally left untouched: the solution of third and fourth-degree equations.

An important example of Renaissance mathematics is the geometry of perspective. The major figures at this time were Leone Battista Alberti (1404–1472), Piero della Francesca (c. 1410–1492), Leonardo (1452–1519), and Albrecht Durer (1471–1528). The mathematics was developed expressly for painters representing the three-dimensional world on a two-dimensional canvas. The ideas were new and the practitioners were generally ignorant of Greek authors.

Baroque Mathematics. Grout characterizes the "Baroque" in terms of a new means of expression. He writes: "Just as seventeenth-century philosophers were discarding outmoded ways of thinking about the world and establishing other more fruitful rationales, the contemporary musicians were seeking out other realms of the emotions and an expanded language in which to cope with the new needs of expression."

There is, at this time, a corresponding new mathematical language: algebra. From the earliest traces of mathematics in ancient civilizations, the common element had been geometry. Even arithmetic, highly developed by the Greeks, was for them firmly embedded in geometry. The great discovery by Descartes (1596–1650) and Fermat (1601–1665) was that geometric forms and ideas could be expressed algebraically. Their method, analytic geometry, was the engine that enabled mathematics to make great leaps forward.

Grout also sees in the Baroque the somewhat incompatible tendencies of emotional intensity and precision. "Baroque music shows conflict and tension between the centrifugal forces of freedom of expression and the centripetal forces of discipline and order in a musical composition. This tension, always latent in any work of art, was eventually made overt and consciously exploited by Baroque musicians; and this acknowledged dualism is the most important single principle which distinguishes between the music of this period and that of the Renaissance."

We can see this dualism in mathematics. The duality is between algebra and geometry. Geometry was the disciplined side of mathematics. It was connected directly to the masterpieces of Greek science, works of rigor and power which were unmatched in the sixteenth century. In algebra was the "freedom of expression," especially the enthusiastic and almost careless use of infinites and infinitesimals, concepts explicitly forbidden by the Greeks.

The tension was very real. In the hands of intuitive geniuses such as Leibniz, the Bernoulis, and later, Euler (1707–1783), algebraic techniques produced a wealth of fundamental results. With others, however, contradictions and falsehoods beckoned alarmingly on every side. The gap between the certainty of geometry and the apparent lack of substance in algebra worried and divided mathematicians and philosophers. Thomas Hobbes (1588–1679) felt the excesses of algebra were totally unjustified. He described Wallis' *Arithmetica Infinitorum* as "a scab of symbols." The most famous attack was by Bishop Berkeley (1685–1753) who savagely mocked imprecision in the work of Newton and others.

Classical Mathematics. Here is Grout on Classical music: "The ideal music of the middle and later eighteenth century, then, might be described as follows: its language should be universal, not limited by national boundaries; it should be noble as well as entertaining; it should be expressive within the bounds of decorum; it should be 'natural', in the sense of being free of needless technical complications and capable of immediately pleasing any normally sensitive listener."

The music of this period spoke immediately to its listeners. So too must the mathematics we call "Classical." Rather than concerning itself with theory, with philosophy, it must be well-motivated by the real world. What Grout and others mean in discussing Classical art, is that the artists communicate easily and directly, that their work is understood and appreciated at once. Compare this to Kline: "Far more than in any other century, the mathematical work of the eighteenth was directly inspired by physical problems. In fact, one can say that the goal of the work was not mathematics, but rather the solution of physical problems; mathematics was just a means to physical ends."

The tension mentioned earlier remained throughout this era, but at a low level. Mathematicians "dared merely to apply the rules and yet assert the reliability of their conclusions." [Kline]. The justification was that it worked; worked in the sense that when applied to physical problems, the new analysis provided answers that were physically verified. Thus, mathematics was "natural" in the sense that it was grounded in the natural world. "The physical meaning of the mathematics guided the mathematical steps and often supplied partial arguments to fill in nonmathematical steps."

Mathematicians of this era sensed there were formal inadequacies in their methods, but on the whole, did not consider them a problem. The employment of mathematics, like the enjoyment of music, tended to be "free of needless technical complications."

Romantic Mathematics. "Romantic" is seemingly the most incongruous adjective we could apply to mathematics. Public impressions of mathematics are far removed from romance, and yet the term can be most apt. Here is Grout describing some of the chief characteristics of the romantic: "... romantic art differs from classic art by its greater emphasis on the qualities of remoteness and strangeness...," "Another fundamental trait of romanticism is boundlessness...," "romantic art aspires... to seize eternity...," "... Romanticism cherishes freedom, movement,

passion, and endless pursuit of the unattainable. Just because its goal can never be attained, romantic art is haunted by a spirit of longing, of yearning after an impossible fulfillment."

Two very mathematical, very modern ideas are expressed here, the concept of the *infinite*, and the concept of the *impossible*. These embody the essence of what I call "romantic mathematics." Both flowered in the nineteenth century.

Let's take infinity first. For thousands of years, the human race had alternately shunned and flirted with the absolute infinite. It was rejected by Aristotle, relished by Lucretius, embraced by Bruno, discarded by Galileo, etc. etc., never rising above philosophy or religion. The first tentative steps to mathematical status were taken by Augustin Cauchy early in the nineteenth century. By 1900, Georg Cantor had lain the foundation for a theory of infinite numbers that would have scandalized the ancient world.

More dramatic is the story of impossibility. In the early nineteenth century, a series of problems which had plagued and motivated mathematicians from ancient times were found to have no solution. Three famous examples come to mind: the trisection of angles by straight-edge and compass, the representation of pi with radicals ("squaring the circle"), and the proof of Euclid's fifth postulate. All were found to be impossible.

Let's look more closely at the last example. Euclid's *Elements* was a compilation and distillation of the known mathematical world. It was a self-contained system that began with five postulates, statements that were accepted as obvious and not requiring proof. Almost from the start, however, there were questions about one of them: the fifth. This statement lacked the simplicity and compelling nature of the first four. Later authors felt that it could and should be proved from the others. For two thousand years, the problem festered. At times, it was put aside, at times work was intense. At times, a solution seemed near. Several writers, indeed, felt they had proved the fifth postulate. Toward the end of the 18th century, the possibility began to be considered that the task was *impossible*.

This was a giant step to take, and one which seemed in open rebellion with the Enlightenment. From our perspective, it is understandable that this would be problematic before the Romantic era. Indeed, it is possible that Gauss, the most powerful and respected mathematician of his age, had come to the conclusion that the problem was impossible over forty years before publishing.

Consider now the logical implications. If one believes that the fifth postulate can't be proved, then one must imagine the existence of a geometry in which the first four axioms are true and fifth false. Otherwise, the acceptance of the first four *compels the acceptance of the fifth*. If we can't prove the fifth then there must exist a geometry unlike any previously known. Here is the important point: mathematicians could accept that the axiom was not provable, yet they could not make the logical step and imagine a different geometry. This step was taken early in the nineteenth century.

The existence of new and strange worlds: *romantic mathematics*. The importance of this particular example is seen in the fact that while all the pieces of the puzzle were in the hands of mathematicians certainly for hundreds of years and arguably for thousands, it remained unsolved. When it *was* solved, it was solved independently by at least three, possibly five researchers, Gauss (perhaps around 1813), Bolyai (c. 1823), Lobachevsky (1827), Schweikart (c. 1812), and Young (Canada, 1860). *Why?*

The immediate suspicion is that something besides mathematics was at work. Our explanation is that the very environment, the intellectual climate of nine-

teenth century romanticism is the reason. Einstein said "Imagination is more important than knowledge." Indeed, the world had all the requisite knowledge, but until the nineteenth century, it did not have the requisite imagination.

After Romanticism. We should end with the romantic era. We are still too close to the twentieth century to understand it, especially when we are trying to grasp a phenomenon as delicate as the artistic milieu of mathematics. There are many identifiable movements: Post-Romanticism, Neo-Classicism, Impressionism, atonal music, and jazz. It is not clear that strong parallels should be drawn before we have the perspective to see what was really going on. Nonetheless, I will make a few irresponsible guesses at the end of this paper.

3. LOOKING AT CORRESPONDING PERIODS. I have set our discussion of mathematical history in a musical context, and so the reader may be unimpressed if the chronologies fit. I claim, nonetheless, that the correspondence is non-trivial. My definition of each mathematical period is inspired by the attributes of the musical one, not by the time frame. For example, I examined Grout's characteristics of the Baroque to decide what mathematics belongs to the Baroque.

In any case, the fit is very good. It is especially good when compared to the chronologies of painting and literature. Running through intellectual history is a significant time lag. Musical periods are generally late. They seem to appear noticeably later than the corresponding periods in the other arts. As we shall see, mathematics appears to match this gap.

Comparing Renaissances. Grout places the Renaissance in music in the range 1450–1600. I place the Renaissance in mathematics at 1500–1600. Gardner puts the Renaissance in art in the fifteenth and sixteenth centuries, with a "proto-Renaissance beginning as early as the late thirteenth century." Benet marks the Renaissance "from the mid-14th century to the end of the 16th century," although it may be referring to a general Renaissance not particularized to literature.

The fit is not merely chronological. Consider the following from Kline: "The Renaissance did not produce any brilliant new results in mathematics. The minor progress in this area contrasts with the achievements in literature, painting, and architecture, where masterpieces that still form part of our culture were created ... " *Kline does not mention music.* In fact, the Renaissance did produce artists and writers whose works remain important and significant today. Consider Dante, Petrarch, Boccaçio, and Shakespeare, Giotto, Botticelli, Michelangelo, Durer, and Leonardo. At the same time, there are almost no comparable figures in music. Only Monteverdi has much popularity today. The names of Palestrina, Praetorius, di Lasso, Byrd, and Gabrielli are known, perhaps, but not to the general public, still less, their work.

Comparing Baroque Periods. Grout places the Baroque loosely between 1600 and 1750. "Baroque" does not seem to be recognized as a literary period. Close in feeling, however, are the metaphysical poets, Donne, Crawshaw, Marvell, also such writers as John Bunyan, all 17th century. Painting and architecture are roughly similar.

Analytic geometry, was developed around 1630. The crowning achievement of the mathematical Baroque, however, was the calculus. Many of the pieces had lain at hand for thousands of years. The critical ideas, however, could hardly be expressed, let alone discovered, without the language of analytic geometry. Not

surprisingly, these ideas were discovered independently by the two greatest minds of the age, Newton (1642–1727) and Leibniz (1646–1716) soon after that language had been established.

One could argue that the Baroque in music and mathematics was not "mature" until the second half of the seventeenth century. In art, maturity was reached much earlier with such figures as Caravaggio (1573–1610), Velasquez (1599–1660), Rubens (1577–1640), Hals (c. 1580–1666), Poussin (1594–1665), and Rembrandt (1606–69).

Comparing Classical Periods. Grout sets the Classical period in music from 1770 to somewhere between 1800 and 1830. Gardner makes no reference to the Classical in art, but the Rococo might be considered equivalent. Indeed, the painter most often compared with Mozart (1756–1791) in spirit and feeling is Watteau (1684–1721). Gardner begins the Rococo early in the eighteenth century.

Literature is more difficult to place, but probably its Classical period is significantly earlier than music's. It might begin with figures such as Dryden (late 17th c). Benet describes Dryden's first works as essentially Baroque: "extravagant late metaphysical," and his later works as Classical: "restrained and natural," "orderly, lucid."

Mathematicians characteristic of the mood and content of the Classical period are Euler (1707–1783), Lagrange (1736–1813), Legendre (1752–1833), and Laplace (1749–1827). This seems to fit better with music than with art.

Consider also the subsidence of the Baroque tension. The great debate over the foundations of calculus, over the infinite, over the place of algebra simply faded away. The issues had in no way been resolved. It has been something of a mystery why the argument died down. In the light of our analysis, however, it does not seem strange. If mathematics is indeed like music, it is natural that mathematicians might simply move on. Grout quotes J. J. Quantz, a critic, writing in 1752, "... the old composers were too much absorbed with musical 'tricks' and carried them too far, so that they neglected the essential thing in music, which is to move and please." On the other hand, the debate raged in the first half of the eighteenth century—further evidence that the Classical period in mathematics begins late in the century, as with music.

Comparing Romantic Periods. The Romantic period is an excellent case for the consanguinity of mathematics and music. Gardner dates the first stirrings of Romantic art early in the eighteenth century in English Neo-Gothic. In Europe, it takes hold as Neo-classicism by, say, 1750. Romanticism in literature is also early, late eighteenth century. Rousseau (1712–1778) is sometimes called the father of romanticism. The first gothic novels appeared in the late eighteenth century.

The Romantic era in both mathematics and music did not begin until early in the nineteenth century. I consider the case for affinity here is especially strong because the identification of romantic mathematics is especially apt.

Summarizing (abbreviations indicate roughly the middle of the period):

	1400		*1500*		*1600*		*1700*		*1800*		*1900*
Music				Ren			Bar		Cla	Rom	
Mathematics					Ren		Bar		Cla	Rom	
Art		Ren				Bar		Cla	Rom		
Literature		Ren				Bar		Cla	Rom		

4. LEAPS, STUMBLES, AND ANTECEDENTS

Am I being fair? My definitions of "Renaissance mathematics," "Baroque mathematics," etc. were not derived in complete ignorance of the relative histories. This was unavoidable. It is certainly conceivable that I was corrupted in the process.

On the other hand, I am not making extravagant statements. I am only calling attention to parallels in the development of mathematics and music. These are certainly there. I don't claim there aren't others as well. Intellectual history is not a science. Perhaps others should attempt to chart similar analogies, between, say, painting and mathematics, or between music and physics. Such a project might well be successful without invalidating this study, however. As I remarked earlier, mathematics is "like" many fields, and in different ways. A new approach could simply reveal a correspondence of an entirely new sort.

Am I being simple-minded? Most of our definitions are based on descriptions from the third edition of Grout's *A History of Western Music*, a popular text first published in 1960. It's not a profound work, and it's probably dated.

Simple-minded yes, but appropriately so. Our focus is on the cultural context of mathematics and music. Mainstream and contemporaneous works such as Grout, Gardner and Kline provide excellent vantage points. In fact, the study could have been performed using texts a hundred years old, as long as they were largely reflective of the prevailing views.

Haven't I missed something? I have mapped the analogy in detail, but I haven't said a word about the mechanism. If music and mathematics develop in parallel, how and why does this happen?

I don't know. That may be difficult to discover. The effect I claim to have noticed here is delicate in the extreme. It is only by looking at the sweep of hundreds of years that I feel confident there is something there at all.

Am I being parochial? I've looked exclusively at western culture. Isn't it possible the conclusions do not carry over to other societies?

I deliberately restricted attention to areas in which I felt I had some knowledge. It may be that the phenomena I observed are not present in other cultural traditions. I would be most interested in studies which explored this.

And anyhow, is this new? Surely others have considered the cultural context of mathematics.

There are antecedents. First of all, there are works which locate mathematics in the stream of intellectual history. Kline's *Mathematics in Western Culture* is a good example. Such studies tend to treat mathematics as a science, however, and concentrate on its relationship with physics, chemistry, etc.

Many writers discuss mathematics and music. The approach is usually philosophical; the analogies between the disciplines as they have actually developed are not noticed. An exception is Tymoczko's paper, "Value Judgements in Mathematics: Can We Treat Mathematics as an Art?" which explores the two as they are practiced, asking such questions as: what corresponds in mathematics to musical performance? In recent work Tymoczko has pioneered and promoted the conception of mathematics as a cultural activity.

Especially significant is the work of Yves Hellegouarch who has written several papers on mathematics and music. Many treat music mathematically, but a few

move in the direction of this study. Most intriguing is "Le Romantisme des les mathematiques" which discusses the romantic features of mathematics, especially in the nineteenth century, and draws parallels to other artistic movements. Prof. Hellegouarch has also linked mathematics and music in education.

There are works which discuss the Romantic movement and the sciences. I found Knight's "Romanticism and the Sciences" intriguing. It focuses on the life of Humphry Davy, poet and chemist using his life span, 1778–1829 to date the period. This seems in serious disagreement with the analysis in this paper, which starts romanticism in mathematics much later. There is, however, *no mention* of mathematics in the paper, or for that matter, in the volume of papers in which it appears. The impulse to separate mathematics from the physical sciences (quite common in the history of science) is significant. It seems likely to me that science, or physics for example, correlates better with literature than with music (or with mathematics).

There have been studies which attempt to discern cultural differences among the mathematics of various countries. This sort of work is called 'ethnomathematics'. Though it is not quite what we are doing here, it is predicated on similar principles.

Jamie James' recent book, *The Music of the Spheres*, deserves special mention. He traces the history in our culture of the conception of the "musical universe," the belief that everything in our world has an explanation, a purpose, a meaning, and a relationship with everything else. It is a theme at least as old as Pythagoras in which "music" and "harmony" have a larger meaning and unite the physical universe with the spiritual. The effect of this theme on music, however, is the chief focus, and music is related more to science than mathematics.

Finally, I note Scott Buchanan's *Poetry and Mathematics*. Buchanan develops a rather ideosyncratic view of mathematics, but makes a strong case for its similarity to poetry. The emphasis on economy of expression is a particularly telling point.

5. GUESSES

Atonal mathematics: This century swirls, from the present perspective, in turbulent musical and mathematical currents. Nonetheless, a reasonably strong argument can be made linking the atonal movement in music with the drive towards formalism in mathematics. Russell and Whitehead's *Principia Mathematica*, for example, the work of Frege, Peano, and Zermelo, mark a radical departure in mathematics. Foundations are now studied for their own sake. In both mathematics and music, it was formerly the case that the work preceded the theory. First Mozart, then Kochel. First Leibniz, then Cauchy. Now it is reversed: first, twelve-tone theory, then Shoenberg. First Zermelo-Fraenkel Set theory, then large cardinals.

Mathematics Jazz: Musically, jazz is the product of two entirely different musical traditions. It is characterized by standard ideas, chords, rhythms, instruments, combined and used in radical and creative ways. A mathematical candidate for jazz is topology. Topology brings together geometry, analysis, algebra, and combinatorics in a field that is totally new. Coincidentally, the jazz player's careless (to the classical musician) treatment of *meter* corresponds neatly to the topologist's disregard of *metric*.

6. IRRESPONSIBLE GUESSES. What follows now is *highly* speculative. As noted earlier, the phenomenon we study is most subtle. Any attempt to discern it in recent movements, isolated cultures, or individuals is extremely risky!

Minimalist Mathematics: In the last 15 years or so, a field of mathematics has arisen that concerns itself with finding the minimum assumptions needed to prove various theorems. It is called "reverse mathematics." Within mathematics, it has not achieved wide recognition, though it is known to most atonal (formal) mathematicians.

New Age Mathematics: Fractal geometry certainly seems in the right mood, with the gauzy pictures it produces of imaginary landscapes. On the other hand, the related subjects of chaos theory and catastrophe theory may be more appropriate. These fields model life qualitatively, not quantitatively, recognizing that most aspects of the world cannot be understood in terms of precise numerical structures.

Punk Mathematics: So-called "brute force" techniques might be described as "punk." So also might the recent practice of using computers to assist in proofs, and to gather inductive evidence when a deductive proof is out of reach.

Here is another: the gradual expansion of what we call "numbers" is an important thread in the history of mathematics. In recent years, however, the operative number system consists merely of those numbers expressible in so many bits by a computer. This a finite system containing no square roots, no irrational numbers, not even *pi*. In reality, most such systems do not even contain the fraction 1/3, just an approximation, .333333. Punk?

Numbers: Number lies at the very heart of mathematics. The history of number is a thread that runs through the history of mathematics from the very earliest times. Is there a possible correspondence between number and musical instruments? A case in point might be the development of keyboard instruments.

Personalities: For the reasons discussed earlier, it is probably *not* a good idea to look for the musical Gauss or the mathematical Beethoven, though it is certainly tempting (especially in the case of Gauss and Beethoven). Chance must definitely play a role in the development of discipline, and that role is magnified in personalities.

On the other hand, there are a few persons who have operated in several fields. I was tempted to look for lessons from this, but on reflection it does not seem justified. Rousseau, for example, composed music. The "father of romanticism" wrote pieces that are thoroughly Baroque. Is this further evidence that musical periods lag behind literary ones? Not necessarily. One could argue that he merely wrote and composed in the style of the times. I have labelled one style "Romantic" and the other "Baroque," but it is the labelling we are debating. If I had labelled that sort of music "Romantic," would the example of Rousseau support a claim that musical periods coincide with literary ones?

For what it is worth, the great English architect Christopher Wren (1632–1723) was also a mathematician. He is described as a transitional figure, his buildings having characteristics of the Baroque and the Classical. They seem more Classical to me. In any case, his mathematics was almost pre-Baroque.

7. CONCLUSIONS. I have presented, I believe, a *prima facie* case for the parallel development of mathematics and music. It embodies a very different understanding of what is popularly imagined a "scientific" discipline. It suggests that the dynamics of mathematical invention resemble the dynamics of artistic invention.

Many of us in mathematics have long felt our field is misunderstood. Children are taught from an early age that mathematics is exact, unforgiving, and arbitrary. Their relationship to it is complex and troubled. They see it as a trap, a contest in which they must guess what the teacher is thinking. For most, it is alternately boring and painful, a game they cannot win, a puzzle where the solution is rarely complete. The situation is perpetuated by the large number of teachers who share, sometimes consciously, sometimes unconsciously, this grim view of mathematics.

It's a terrible way to educate children. Few can ever become fond of a subject treated this way. The import of this study is that it is worse than bad strategy, it is a fundamentally flawed picture of mathematics. I would argue that mathematics should be taught as music is taught. Students should make mathematics *together* (as in fact professional mathematicians do), not alone. Creative ideas should be stressed over the "right way to do it." Mathematics should be a treat, not a chore. And finally, students should perform mathematics; they should *sing* mathematics and *dance* mathematics.

A lovely example of what can be accomplished in this way is Zoltan Dienes' *Mathematics Through the Senses*, unfortunately out of print. Dienes presents an astonishing variety of settings for exploring significant mathematics at the grade-school level. The very first, the "three cornered waltz" is a perfect example. The basic dance form is presented which grade-school children, in groups of three, can perform and vary. The steps are simple, but they lead the young dancers into an exploration of the symmetric group S_3.

As an art, mathematics has not had a wide audience. Its passions and its pleasures are denied to most. To the public, it is cold and barren. To poets, it is cold and austere. Both are quite wrong. Mathematics is not divine; it is mortal. Mathematics is not law; it is taste. Mathematics is not calculation, but communication. The best mathematics is not *true*; it is *beautiful*.

REFERENCES

Benet, William Rose, *The Reader's Encyclopedia*, New York: Thomas Y. Crowell Co., 1965.

Buchanan, Scott, *Poetry and Mathematics*, John Day Co. 1929.

Dienes, Zoltan, *Mathematics Through the Senses, Games, Dance, and Art*, NFER Pub. Co. Ltd., 1973.

Gardner, Helen, *Art Through the Ages*, 7th edition, revised by Hosrt De La Croix and Richard G. Tansey, New York: Harcourt, Brace, 1980.

Grout, Donald Jay, *History of Music*, 3rd edition, New York: W. W. Norton & Co., 1980.

Hellegouarch, Yves, "Le Romantisme dans les Mathematiques, ou un Regard Oblique sur les Mathematiques du XIX^e Siecle," Yves Hellegouarch, in *Destin del l'Art Desseins de la Science*, Bessot, Hellegouarch, le Goff, eds.

James, Jamie, *The Music of the Spheres*, Grove Press, 1993.

Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Oxford: Oxford University Press, 1972.

Kline, Morris, *Mathematics in Western Culture*, Oxford: Oxford University Press, 1953.
Knight, David, "Romanticism and the Sciences," in *Romanticism and the Sciences*, Andrew Cunningham, ed., Cambridge: Cambridge University Press, 1990.
Tymoczko, Thomas, "Humanistic and Utilitarian Aspects of Mathematics," to appear in the *Proceedings of the 7th International Congress of Mathematical Education.*
Tymoczko, Thomas, "Value Judgements in Mathematics: Can We Treat Mathematics As An Art?", to appear in *Essays in Humanistic Mathematics*, Alvin White, ed., MAA Monographs.

Clark Science Center
Smith College
Northampton, MA 01063
jhenle@smith.edu.

PICTURE PUZZLE
(from the collection of Paul Halmos)

This picture was taken almost a quarter of a century ago, and he still hasn't learned to spell his name "right".
(see page 50)

4. G. H. Hardy, E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford 1959.
5. P. Ribenboim, *The Book of Prime Number Records*, Springer-Verlag 1989.

D. Borwein:
Department of Mathematics
University of Western Ontario
London, Ontario
N6A 5B7, Canada
dborwein@uwo.ca

J. M. Borwein:
Department of Mathematics and Statistics
Simon Fraser University
Burnaby BC V5A 1S6, Canada
jborwein@cecm.sfu.ca

P. Borwein:
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
pborwein@cecm.sfu.ca

R. Girgensohn:
Institut für Mathematik
Medizinische Universität Lübeck
P-23560 Lübeck
GERMANY
girgenso@informatik.mu-luebeck.de

Rounding Error?

From article #155 about rocks in Antarctica in the book *365 Surprising Scientific Facts, Breakthroughs, and Discoveries* by S. B. McGrayne, Wiley, 1994:

...north-facing rocks that soak up polar sunlight can be 15°C (60°F) warmer than the surrounding air.

Submitted by H. Turner Laquer
Idaho State University

Answer to Picture Puzzle

(p. 29)

Jack Schwartz.

Tantrices of Spherical Curves

Bruce Solomon

1. INTRODUCTION

(1.1) If the speed of a smooth closed curve σ on the unit sphere $\mathbf{S}^2$ never vanishes, than σ's *normalized* velocity vector $\tau := \dot{\sigma}/|\dot{\sigma}|$ sweeps out a new closed curve on $\mathbf{S}^2$, often called the *tangent indicatrix* of σ. Here we shall simply call τ the *tantrix* of σ.

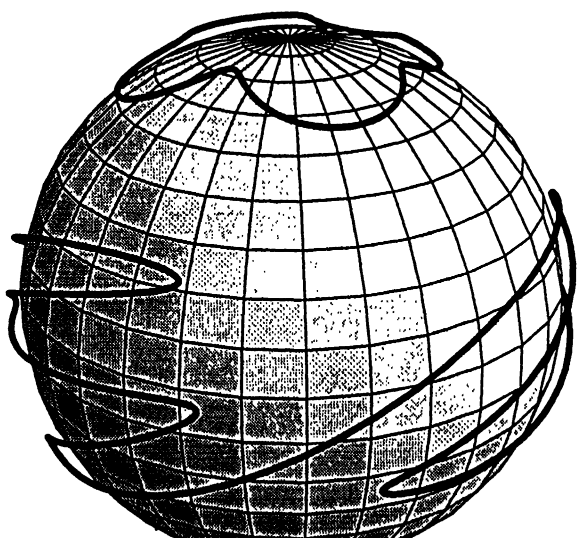


Figure 1. A closed curve σ on $\mathbf{S}^2$ (shorter loop near north pole) and its tantrix.

While every loop σ on $\mathbf{S}^2$ with non-vanishing speed defines a "tantricial" loop in this way, *not every loop is tantricial.* Here, we will completely expose the non-obvious—but lovely—obstruction to this converse.

To begin, note that *if the speed of σ never vanishes, neither will that of τ.* In fact, the speed of τ, computed relative to arclength along σ, gives the *curvature* of σ as a curve in $\mathbf{R}^3$. But σ, lying on $\mathbf{S}^2$, can nowhere approximate a straight line to second order: its curvature—the speed of τ—will never vanish. Referring non-experts to the sidebar discussions of "immersion" and "arclength" for further details, we conclude more precisely that *if σ immerses the circle in $\mathbf{S}^2$, so will τ.*

This observation provokes our main question:
When does one immersed circle on $\mathbf{S}^2$ form the tantrix of another?

Immersion. Roughly speaking, a curve in $\mathbf{S}^2$ constitutes an *immersion* if it has no corners or cusps, though it may cross itself. Analytically, one guarantees these properties by insisting that *the curve's velocity never-vanishes.* A mapping σ: $\mathbf{S}^1 \to \mathbf{S}^2$ thus forms an *immersion of the circle into the sphere* iff for all $\theta \in \mathbf{R}$, we have

$$\left|\frac{d}{d\theta}\sigma(e^{i\theta})\right| > 0$$

In this article, we shall deal only with twice continuously differentiable ("smooth") immersions.

Arclength Parametrization. A basic fact from the differential geometry of curves states that one can reparametrize any immersion using *arclength.* This means we can give the curve *unit-speed* relative to some new parameter s. In the context above, for example, we can arrange

$$\left|\frac{d}{ds}\sigma(e^{i\theta(s)})\right| \equiv 1$$

The tantrix τ of σ will now *equal* σ's velocity vector: $\tau = d\sigma/ds$. As claimed in our introduction, τ will then *immerse* the circle into $\mathbf{S}^2$—$d\tau/ds$ will never vanish. In fact, we can easily prove $|d\tau/ds| \geq 1$:

$$0 = \frac{1}{2}\frac{d^2}{ds^2}(|\sigma|^2) = \left|\frac{d\sigma}{ds}\right|^2 + \frac{d^2\sigma}{ds^2}\cdot\sigma = 1 + \frac{d\tau}{ds}\cdot\sigma,$$

hence (by Cauchy-Schwartz)

$$(*) \qquad \left|\frac{d\tau}{ds}\right| = \left|\frac{d\tau}{ds}\right||\sigma| \geq \left|\frac{d\tau}{ds}\cdot\sigma\right| = 1.$$

Since $|d^2\sigma/ds^2|$ gives the *curvature* of a unit-speed curve σ, and we have $d\tau/ds = d^2\sigma/ds^2$, this shows too that *any curve on* $\mathbf{S}^2$ *has curvature at least 1.*

Joel Weiner recently discovered the answer "by accident" while working on the seemingly unrelated topic of flat tori in $\mathbf{S}^3$:

(1.2) Theorem. ([W]) *An immersed circle τ in $\mathbf{S}^2$ forms a tantrix if and only if it has total geodesic curvature zero, and contains no subarc with total geodesic curvature π.*

While Weiner derives this simple fact as a corollary to his results on flat tori, he remarks that "it would be nice to have a curve-theoretic proof" of the result. Here we provide such a proof, and present some related facts. For one, we have

(1.3) Theorem. *An immersed circle in $\mathbf{S}^2$ and its tantrix share a regular homotopy class. A tantrix in the equator's class always bounds oriented area 2π (mod 4π). A tantrix in the other class bounds area zero.*

As we discuss in a sidebar (§5), *regular homotopy* means "homotopy through immersions," and in $\mathbf{S}^2$, immersed circles come in just two regular homotopy flavors, represented by single and double traversals of the equator respectively. Notice that these two curves each form their own tantrices, and indeed, bound areas 2π and $0 \pmod{4\pi}$ respectively. (See §5 for the notion of "oriented area $\bmod 4\pi$.")

The claim about area in this Theorem suggests a connection with the classical result known as "Jacobi's Theorem for space curves," which appears in standard texts such as Chern [**Ch**, p. 44], Spivak [**Sp**, §6.12], and DoCarmo [**DoC**]. In fact, Jacobi's theorem follows immediately from the earlier half of Weiner's theorem, which we can state like this:

Proposition 3.1. *The tantrix of an immersed circle in* $\mathbf{S}^2$ *always has total geodesic curvature zero, and if non-self-intersecting, bounds area* 2π.

We postpone the elementary proof of this fact to §3, but we show right now how to deduce Jacobi's Theorem from it:

(1.4) Corollary. ("Jacobi's Theorem") *The principal normal indicatrix of a closed space curve with non-vanishing curvature, if embedded,* bisects *the area of* $\mathbf{S}^2$.

(1.5) *Proof.* The unit tangent vector T along a closed space curve γ maps $\mathbf{S}^1$ into $\mathbf{S}^2$, and here, *immerses* $\mathbf{S}^1$, since the curvature κ of γ never vanishes. Indeed, when s represents arclength along γ, the Frenet formula $dT/ds = \kappa N$, shows that the tantrix of T coincides with the *normal* indicatrix N of γ. If N never crosses itself, it bounds a domain $\Omega \subset \mathbf{S}^2$, with area $0 < |\Omega| < 4\pi$. Proposition 3.1 then forces $|\Omega| = 2\pi = \frac{1}{2}|\mathbf{S}^2|$. Q.E.D.

Note that Jacobi's Theorem does *not*, conversely, imply Proposition 3.1. For, the typical closed curve on $\mathbf{S}^2$ does *not* form the tangent indicatrix of a closed space curve; the latter always have length at least 2π, (see [**Ch**, §4], for instance), and can never lie in an open hemisphere (exercise). *We therefore submit that Proposition 3.1—not Jacobi's theorem—would better serve texts like* [**Ch**], [**Sp**], *and* [**DoC**]. With the aim of emphasizing this possibility, we prove the "if and only if" statement of Theorem 1.2 as separate subsidiary results—Propositions 3.1 and 4.1—and prove them by completely elementary techniques.

To deal with the subtler notions of oriented area and regular homotopy invoked by Theorem 1.3, we must resort to more sophisticated means in §5. We have tried to write §5 so that novice readers can follow its main points, if not its details, however, so the earlier sections borrow nothing from this final one.

Incidentally, Theorem 1.3 implies an amusing corollary:

(1.6) Corollary. *Any immersed circle regularly homotopic to a figure-eight in* $\mathbf{S}^2$ *must cross some great circle orthogonally—and in the same direction—at least twice.*

Proof: One can regularly homotope any figure-eight in $\mathbf{S}^2$ to a double-equator, and the *tantrix* of such a curve must then self-intersect, since the Proposition now excludes it from the regular homotopy class of the single-equator. A self-crossing of the tantrix, however, signals two or more points on the original closed curve with the same oriented tangent vector. The corollary follows directly, because the set of points on $\mathbf{S}^2$ that share a particular tangent direction form the great circle perpendicular to that direction. Q.E.D.

ACKNOWLEDGMENT. The author developed these ideas at the Institut des Hautes Etudes Scientifiques in Bures-sur-Yvette (France), which he gratefully acknowledges for its excellent hospitality and much-appreciated support. Thanks also to Joe Fu for his interest and remarks, and to the referees for their helpful comments.

2. CURVES ON $\mathbf{S}^2$ AND A KEY LEMMA. Here we prepare for later arguments with some elementary calculations for curves on $\mathbf{S}^2$, and prove an easy—but very illuminating—lemma.

(2.1) Consider an arbitrary immersed curve on $\mathbf{S}^2$, and call it τ, since we will soon want to ask whether it arises as the tantrix of some other spherical curve σ. Refer τ to an arclength parameter t (we will save "s" for arclength along σ), so that $\dot{\tau} := d\tau/dt$ has length 1. Define also the unit normal ν to τ in $\mathbf{S}^2$ via $\nu := \tau \times \dot{\tau}$, so that $\{\tau, \dot{\tau}, \nu\}$ forms an oriented orthonormal basis for $\mathbf{R}^3$, with $\dot{\tau}$ and ν spanning the tangent plane to $\mathbf{S}^2$ at $\tau(t)$ for each t. By definition, when one expands the *second* derivative $\ddot{\tau}$ in terms of this frame, the coefficient of ν gives the *geodesic curvature* κ_g of τ. To get the two remaining coefficients, we simply compute

$$\ddot{\tau} \cdot \dot{\tau} = \frac{1}{2}\frac{d}{dt}|\dot{\tau}|^2 = 0,$$

> **Geodesic Curvature.** As mentioned above, the acceleration vector of a unit-speed space curve gives (via its length) the *curvature* κ of that curve. One can parametrize any immersed curve by arclength, making this definition quite general. (*Exercise:* For a radius-r circle in the plane, $\kappa = 1/r$.)
> For unit-speed curves on *surfaces*, we get the *geodesic* curvature—κ_g—by measuring the length of the *surface-tangential component* of the acceleration. (*Exercise*: For a radius-r circle on $\mathbf{S}^2$, $\kappa_g = \sqrt{1 - r^2}\,/r$.) As shown in §3, geodesic curvature relates very closely to parallelism. Indeed, by Equation 3.2.1, κ_g gives the rate at which a *parallel* vectorfield rotates relative to the *tangen* vectorfield of a unit-speed curve. (*Exercise*: How much will a Foucault pendulum at latitude ϕ precess in 24 hours?)
> *Geodesics*—curves with $\kappa_g \equiv 0$—play a major role in Differential Geometry because they provide the shortest paths connecting pairs of points on surfaces (and on higher dimensional "surfaces").

and

$$\ddot{\tau} \cdot \tau = \frac{1}{2}\frac{d^2}{dt^2}(|\tau|^2) - |\dot{\tau}|^2 \equiv -1.$$

We thus have

$$\ddot{\tau} = \kappa_g \nu - \tau. \tag{2.1.1}$$

Among other things, this fact implies the spherical "Frenet" formula

$$\begin{aligned}\dot{\nu} &= \frac{d}{dt}(\tau \times \dot{\tau}) = \dot{\tau} \times \dot{\tau} + \tau \times \ddot{\tau} = \tau \times (\kappa_g \nu - \tau) = \kappa_g \tau \times \nu \\ &= -\kappa_g \dot{\tau}\end{aligned} \tag{2.1.2}$$

Next, consider a unit vector field σ tangent to $\mathbf{S}^2$ along τ. Characterize the angle

ϕ between σ and $-\dot{\tau}$ at each time t by the equation

$$\sigma(t) = -\cos\phi(t)\cdot\dot{\tau}(t) + \sin\phi(t)\cdot\nu(t). \tag{2.1.3}$$

Differentiate this equation, and expand in terms of $\dot{\tau}$, ν, and τ, using Equations 2.1.1 and 2.1.2:

$$\begin{aligned}\dot{\sigma} &= \dot{\phi}\sin\phi\dot{\tau} - \cos\phi\ddot{\tau} + \dot{\phi}\cos\phi\nu + \sin\phi\dot{\nu} \\ &= \sin\phi\left(\dot{\phi} - \kappa_g\right)\dot{\tau} + \cos\phi\left(\dot{\phi} - \kappa_g\right)\nu + \cos\phi\tau \\ &= \left(\dot{\phi} - \kappa_g\right)(\sin\phi\dot{\tau} + \cos\phi\nu) + \cos\phi\tau \end{aligned} \tag{2.1.4}$$

We will soon apply these facts, but first we record a lemma. Though very simple, this lemma has a surprise bottom line invoking the key notion of *parallelism* for vectorfields tangent to $\mathbf{S}^2$ along a curve (see sidebar).

Parallel Vectorfields along Curves on $\mathbf{S}^2$. In the *plane*, we call a vectorfield $\mathbf{v}$ *parallel* along a path γ if it has constant length and direction:

$$\mathbf{v}\text{ parallel} \quad \Leftrightarrow \quad \frac{d}{dt}\mathbf{v}(\gamma(t)) \equiv 0 \text{ (plane)}$$

Along a curve in $\mathbf{S}^2$, however, one can't generally find a vectorfield *tangent* to the sphere with this property. So we can't define parallelism by the complete vanishing of a spherical vectorfield's derivative. One can, however, ask the derivative to vanish *tangentially*, thus defining $\mathbf{v}$ as *parallel* along γ in $\mathbf{S}^2$ if its derivative always points *normally* to $\mathbf{S}^2$ along γ. Since each point in $\mathbf{S}^2$ *equals* the normal to $\mathbf{S}^2$ at that point, we then have

$$\mathbf{v}\text{ parallel} \quad \Leftrightarrow \quad \frac{d}{dt}\mathbf{v}(\gamma(t)) = f(t)\gamma(t) \text{ (sphere).}$$

for some scalar function f.

This notion of parallism has many applications in Geometry, and even in Physics. For example, as it repeatedly sweeps through its lowest point, the velocity vector of a Foucault Pendulum traces out a parallel vectorfield along a circle of latitude on the rotating Earth—showing physically that the parallel extension of an initial starting vector along a loop in $\mathbf{S}^2$ (*e.g.*, a circle of latitude in the "Foucault" case) won't generally end up where it started after traversing the loop!

(2.2) Lemma. *If s and t denote oriented arclength parameters along a curve σ immersed in $\mathbf{S}^2$, and along its tantrix τ respectively, then*

$$\dot{\sigma} := \frac{d\sigma}{dt} = \tau\frac{ds}{dt} \quad \textit{and} \quad \frac{ds}{dt} > 0.$$

One can therefore regard any immersed curve σ in $\mathbf{S}^2$ as a parallel *vectorfield tangent to $\mathbf{S}^2$ along its own tantrix.*

Proof: Write t as a function of s by basing an arclength integral at some value s_0 of s:

$$t(S) = \int_{s_0}^{S}\left|\frac{d\tau}{ds}\right| ds.$$

We thus have $t'(s) = |d\tau/ds|$—the speed of τ— which must exceed 1, as discussed in our paper's second paragraph (see also Equation $(*)$ of the sidebar item on *Arclength*). In particular, $t'(s) > 0$. Furthermore, $d\tau/ds = d^2\sigma/ds^2$, so our standing smoothness assumption on σ ensures continuity of t'. The Inverse Function Theorem now makes s a function of t, with

$$\frac{ds}{dt} = \frac{1}{t'(s)} > 0,$$

and

$$\dot{\sigma} = \frac{d\sigma}{ds}\frac{ds}{dt} = \tau\frac{ds}{dt}.$$

Having derived the asserted formulas, we now simply observe that $\tau(s)$ coincides with the normal to $\mathbf{S}^2$ at $\tau(s)$, and $\sigma(s)\cdot\tau(s) = d/ds(\frac{1}{2}|\sigma|^2) \equiv 0$, so $\sigma(s)$ lies *tangent* to $\mathbf{S}^2$ at $\tau(s)$ for each s. Moreover, our expression for $\dot{\sigma}$ leaves it with *no* component tangent to $\mathbf{S}^2$. This makes σ *parallel* along τ. Q.E.D.

3. AN ELEMENTARY ANTECEDENT TO JACOBI'S THEOREM. Before indulging in a more complete analysis of tantrices, we supply the easy Proposition that implies Jacobi's Theorem via the argument given in §1.5. Recall that the integral of κ_g over an entire curve defines that curve's *total geodesic curvature.*

(3.1) Proposition. *The tantrix of an immersed circle in* $\mathbf{S}^2$ *always has total geodesic curvature zero, and if embedded, bounds area* 2π.

(3.2) *Proof.* Let τ denote the tantrix of an immersed circle $\sigma \subset \mathbf{S}^2$. By Lemma 2.2, we may regard σ as a vectorfield tangent to $\mathbf{S}^2$ along τ, so Equation 2.1.4 applies. Comparing the expression given there for $\dot{\sigma}$ with the one given by Lemma 2.2, we immediately deduce

$$\dot{\phi} = \kappa_g. \tag{3.2.1}$$

Moreover, σ closes up, making ϕ L-periodic, L denoting the length of the tantrix. Hence

$$0 = \phi(L) - \phi(0) = \int_0^L \dot{\phi}\,dt = \int_\tau \kappa_g.$$

This proves the Proposition's first claim, and Gauss-Bonnet does the rest:

$$|\Omega| = \int_\Omega 1\,dA = \int_\Omega K\,dA = 2\pi - \int_\tau \kappa_g = 2\pi.$$ Q.E.D.

4. INVERTING A TANTRIX. Implicitly, Lemma 2.2 suggests an elegant way to "invert" a tantrix—*i.e.*, to find an immersed circle $\sigma \subset \mathbf{S}^2$ having a given curve τ as tantrix. Namely, *one should simply parallel-translate a unit vector tangent to* $\mathbf{S}^2$ *around* τ. But even when the total geodesic curvature of τ vanishes—an obvious necessity by Proposition 3.1—this procedure can fail: it may not produce an immersion. For instance, along the equator, the parallel "vertical" vectorfield $(0,0,1)$ traces out only one point—hardly an immersed circle. On the other hand, parallel-translating any *non*-vertical unit tangent around the equator yields a circle of latitude, whose tantrix *does* give back the equator. Proposition 4.1 below will resolve this paradox using the oscillation of a total curvature function.

To define that function, observe that the vanishing of total geodesic curvature on an immersed circle $\tau \subset \mathbf{S}^2$

$$\int_\tau \kappa_g = 0,$$

ensures that κ_g has a continuous antiderivative—call it ϕ_τ—on τ. Geometrically, ϕ_τ measures the angle between $-\dot{\tau}$ and some parallel unit vectorfield along τ; this follows from Equations 2.1.3 and 2.1.4. Of course, τ boasts an entire circle of such vectorfields; we have specified ϕ_τ only up to addition of a constant. But no such ambiguity afflicts the *oscillation* of ϕ_τ, defined via

$$\operatorname{osc} \phi_\tau := \sup \phi_\tau - \inf \phi_\tau .$$

Moreover, Equation 3.2.1 says $\dot{\phi}_\tau = \kappa_g$. So $\operatorname{osc} \phi_\tau$ *measure the maximum total geodesic curvature of any subarc of* τ.

(4.1) Proposition. *On the tantrix τ of an immersed circle σ in $\mathbf{S}^2$, we always have $\operatorname{osc}\phi_\tau < \pi$; i.e., no subarc of τ has total geodesic curvature π. Conversely, any immersed circle τ in $\mathbf{S}^2$ having total geodesic curvature zero and $\operatorname{osc} \phi_\tau < \pi$, forms the tantrix of some other immersed circle σ in $\mathbf{S}^2$.*

Proof: Recall that by Lemma 2.2, σ lies tangent to $\mathbf{S}^2$ along τ, with

$$\dot{\sigma} = \tau \frac{ds}{dt} \quad \text{and} \quad \frac{ds}{dt} > 0.$$

These two facts produce immediate consequences via Equation 2.1.4. Namely, on denoting by ϕ the angle between σ and $-\dot{\tau}$ (as characterized by Equation 2.1.3), 2.1.4 implies

$$\dot{\phi} = \kappa_g \quad \text{and} \quad \cos \phi = \frac{ds}{dt} > 0.$$

The first identity here shows that ϕ antidifferentiates κ_g; we can take $\phi_\tau = \phi$. The second then forces $-\pi/2 < \phi_\tau < \pi/2$, which clearly means $\operatorname{osc} \phi_\tau < \pi$, and proves the Proposition's first statement.

To get the converse, consider an immersed circle τ in $\mathbf{S}^2$ with total geodesic curvature zero—so that κ_g has a continuous antiderivative—and assume $\operatorname{osc} \phi_\tau < \pi$. This latter restriction clearly lets us choose an antiderivative ϕ_τ for κ_g with $-\pi/2 < \phi_\tau < \pi/2$.

Having done so, let t denote arclength along τ, and construct the unit normal $\nu = \dot{\tau} \times \tau$ along τ as in §2 above. If we now *define* a tangent vectorfield σ along τ using Equation 2.1.3 with ϕ replaced by ϕ_τ, so that

$$\sigma(t) = -\cos \phi_\tau(t)\dot{\tau}(t) + \sin \phi_\tau(t)\nu(t),$$

then Equation 2.1.4 immediately forces $\dot{\sigma} = \cos \phi_\tau \tau$, because $\dot{\phi}_\tau - \kappa_g = 0$. Since $\cos \phi_\tau$ never vanishes, this makes σ an immersion, with tantrix τ. Q.E.D.

5. AREAS AND REGULAR HOMOTOPY. We now turn to Theorem 1.3, restated here for the reader's convenience:

Theorem 1.3. *An immersed circle in $\mathbf{S}^2$ and its tantrix share a regular homotopy class. A tantrix in the equator's class always bounds area 2π. A tantrix in the other class bounds area zero.*

Regular Homotopy of Curves. *Homotopy* means "continuous deformation." For instance, if we deform a round circle in the plane into a long, narrow ellipse by stretching along an axis, we create a homotopy joining these two closed curves. A homotopy qualifies as *regular* when the curve begins and remains *immersed* throughout the deformation.

Two curves belong to the same regular homotopy *class* if some regular homotopy joins them—like the circle and ellipse mentioned above. Every immersed circle in the *plane* homotopes regularly to an n-times traversed circle for some unique integer n (add $+1$ for each counterclockwise traversal, -1 for clockwise traversals), or to a figure-eight ($n = 0$). In particular, such curves form infinitely many *different* classes. For instance, a simple clockwise circle won't deform into a counterclockwise one without passing through a "cusp" at some intermediate step:

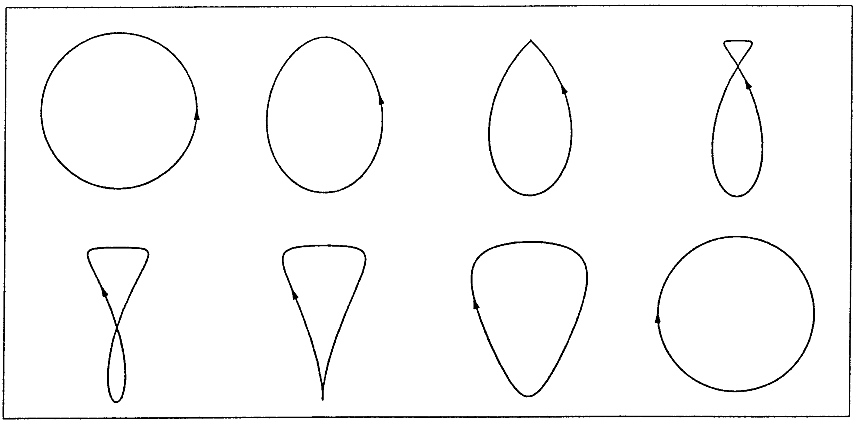


Figure 2. A typical—hence *non*-regular—homotopy from a $+1$ circle to a -1 circle. Like this one, any homotopy between counterclockwise and clockwise circles must pass through a cusp at some stage.

On the *sphere*, however, the situation differs dramatically. For instance, one can homotopically reverse the orientation of a longitudinal circle by simply rotating it 180°, keeping its north and south poles fixed. Moreover, by generalizing this "trick," one can show that on $\mathbf{S}^2$, *immersed circles form only* **two** *regular homotopy classes*. The equator, traced once (or any *odd* number of times) represents one class. Traced twice (or any even number of times), it represents the other class. (*Exercise:* Convince yourself that on $\mathbf{S}^2$, figure-eights belong to the latter class.) For further elaboration of this fascinating topic, see **[P]**.

(5.1) *Areas.* Since the tantrix of an immersed circle can cross itself, thereby cutting the sphere into many sub-domains, the *area* it bounds requires careful definition. One *can* make a good definition, however, because a closed, immersed curve in $\mathbf{S}^2$ always bounds a 2-dimensional *homology class*, and one may assign areas to such classes very naturally.

To see how, let C denote any union of oriented, immersed circles in $\mathbf{S}^2$. Suppose we have a smooth surface S with boundary, and a smooth mapping p: $S \to \mathbf{S}^2$ such that $p(\partial S) = C$. Given C, standard differential topology guarantees

the existence of an S and a p with these attributes: in fact, one can extend any immersion with image C smoothly (through not generally as an immersion) to any smooth surface that spans its domain. Now let ω denote the standard area 2-form on $\mathbf{S}^2$. The prescription

$$\operatorname{area}(C) := \int_S p^*\omega \pmod{4\pi} \tag{5.1.1}$$

makes good sense; it depends on neither S, nor p, since it returns $4\pi \deg(p)$ when $\partial(S) = 0$ (*e.g.*, by the "degree formula" in **[G & P]**). Equation 5.1.1 clearly corroborates all familiar definitions of area for closed curves which *don't* self-intersect in $\mathbf{S}^2$—just choose the subdomain of $\mathbf{S}^2$ bounded by C for S. So we define the area bounded by general closed *immersed* curves using this prescription.

(5.2) *The Unit Tangent Bundle of* $\mathbf{S}^2$. A spanning surface S arises naturally in our proof of Theorem 1.3; it will lie in the 3-dimensional space of all unit vectors tangent to $\mathbf{S}^2$—the sphere's *unit tangent bundle* US^2. We may specify such vectors using their base-points in $\mathbf{S}^2$ and their unit direction vectors—also points in $\mathbf{S}^2$. So we shall regard $\mathbf{US}^2$ as the following subset of $\mathbf{S}^2 \times \mathbf{S}^2$ (in $\mathbf{R}^3 \times \mathbf{R}^3$):

$$\mathbf{US}^2 := \{(\mathbf{u}, \mathbf{v}) \in \mathbf{S}^2 \times \mathbf{S}^2 : \mathbf{u} \cdot \mathbf{v} = 0\}.$$

At each base point $\mathbf{u} \in \mathbf{S}^2$, we find a circle of unit tangent vectors $\mathbf{v}_u$. This makes $\mathbf{US}^2$ a bundle of circles over the 2-sphere. Let p denote the bundle projection that sends each circular fiber to its base-point:

$$p\colon \mathbf{US}^2 \to \mathbf{S}^2, \quad p(\mathbf{v}_u) := \mathbf{u}.$$

The formula

$$(\mathbf{u}, \mathbf{v}) \mapsto (\mathbf{u}, \cos\theta \mathbf{v} + \sin\theta(\mathbf{u} \times \mathbf{v}))$$

rotates all fibers through angle θ (counterclockwise, as viewed from outside the sphere), so by applying $d/d\theta|_{\theta=0}$, we harvest a smooth unit vectorfield on $\mathbf{US}^2$ which flows tangent to the fibers; namely, $(0, \mathbf{u} \times \mathbf{v})$ at the point $(\mathbf{u}, \mathbf{v})$.

Call the one-form dual to this vectorfield α, so that $\alpha(\mathbf{x}) := (0, \mathbf{u} \times \mathbf{v}) \cdot \mathbf{x}$ for any $\mathbf{x}$ tangent to $\mathbf{US}^2$ at $(\mathbf{u}, \mathbf{v})$.† The exterior derivative $d\alpha$, it turns out, gives the *curvature 2-form* on $\mathbf{US}^2$. So since $\mathbf{S}^2$ has constant curvature 1, $d\alpha$ simply pulls back 1 times the area form on $\mathbf{S}^2$, via the projection p:

$$d\alpha = p^*\omega. \tag{5.2.1}$$

Readers unfamiliar with this equation may compute $d\alpha$ directly using the fact that

$$d\alpha(X, Y) = (\nabla_X \alpha)(Y) - (\nabla_Y \alpha)(X) = D_X(0, u \times v) \cdot Y - D_Y(0, u \times v) \cdot X$$

whenever X and Y are tangent to $\mathbf{US}^2$.

Our main argument now comes easily:

(5.3) *Proof of Theorem* 1.3. The tantrix of an immersed circle again *immerses* the circle, as we have noted several times. Note also that multiples of the equator reparametrize their own tantrices. So if we deform a curve σ in $\mathbf{S}^2$ through immersions to some multiple of the equator, its tantrix flows simultaneously to that

†The knowledgeable reader may recognize α as the standard *connection* form on $\mathbf{US}^2$. In particular, if one lifts a smooth unit-speed curve $\tau(t)$ on $\mathbf{S}^2$ to the curve $\gamma = (\tau, \dot\tau)$ in $\mathbf{US}^2$, then $\alpha(\dot\gamma(t))$ returns the geodesic curvature of τ at time t. We shall evaluate α along a *different* lift to prove Theorem 1.3, however.

same multiple equator—and likewise through immersions. In particular, any immersed circle shares a regular homotopy class with its tantrix, as the Theorem's first statement claims.

To get the statements about area, let τ denote the tantrix of an immersed circle σ in $\mathbf{S}^2$, and refer both to an arclength parameter t along τ. Lift τ to the curve $\Lambda_\tau = (\tau, \sigma)$ in $\mathbf{US}^2$. The collinearity of $\dot{\sigma} := d\sigma/dt$ with τ (Lemma 2.2) then makes Λ_τ everywhere *horizontal*—orthogonal to the fibers—in $\mathbf{US}^2$, as an easy calculation shows:

$$\alpha\left(\dot{\Lambda}_\tau\right) = (0, \tau \times \sigma) \cdot (\dot{\tau}, \dot{\sigma}) = (\tau \times \sigma) \cdot \dot{\sigma} \equiv 0. \tag{5.3.1}$$

Now consider some multiple of the equator—call it e: $\mathbf{S}^1 \to \mathbf{S}^2$—and suppose we can deform σ to e through immersions. By lifting the tantrix of each curve in the resulting regular homotopy to $\mathbf{US}^2$ as just described, we smoothly map an annulus S into $\mathbf{US}^2$, with $\partial(p(S)) = p(\Lambda_\tau - \Lambda_e) = \tau - e$. We can therefore calculate the area that $\tau - e$ bounds in $\mathbf{S}^2$ by applying Stokes' Theorem in conjunction with our area prescription (5.1.1), the formula for $d\alpha$ (5.2.1), and the horizontality of lifted tantrices (5.3.1):

$$\operatorname{area}(\tau - e) = \int_S p^*\omega = \int_S d\alpha = \int_{\partial S} \alpha = \int_{\Lambda_\tau - \Lambda_e} 0 = 0.$$

It follows immediately that τ and e bound the same area in $\mathbf{S}^2$. Since the Theorem's claims about area clearly hold for multiples of the equator, hence for e, we now get the same for the arbitrary tantrix τ. Q.E.D.

REFERENCES

[Ch] Chern, S. S., *Curves and Surfaces in Euclidean Space*, **Studies in Global Geometry and Analysis**, Mathematical Association of America, Prentice-Hall Inc., Englewood Cliffs, NJ, 1967.

[DoC] DoCarmo, M. P., **Differential Geometry of Curves and Surfaces**, Prentice-Hall Inc., Englewood Cliffs, NJ, 1976.

[G & P] Guillemin, V. & Pollack, A., **Differential Topology**, Prentice-Hall Inc., Englewood Cliffs, NJ, 1974.

[M] Murphy, C., *"To Be" in Their Bonnets*, Atlantic Monthly, Vol. 270, No. 2 (February 1992). The Atlantic Monthly Co., 745 Boylston St., Boston, MA 02116.

[P] Phillips, A., *Turning a surface inside out*, Scientific American **214**, *May* 1966, *No.* 5, 112–120.

[Sp] Spivak, M., **A Comprehensive Introduction to Differential Geometry, Volume III**, Publish or Perish, Boston, MA, 1975.

[W] Weiner, J., *Flat tori in* $\mathbf{S}^3$ *and their Gauss maps*, Proc. London Math. Soc. (3) **62**, 1991, No. 1, 54–76.

Department of Mathematics
Indiana University
Bloomington, IN 47405
solomon@ucs.indiana.edu

Giuga's Conjecture on Primality

D. Borwein, J. M. Borwein, P. B. Borwein, R. Girgensohn

1. INTRODUCTION. In 1950, G. Giuga formulated the following conjecture ([3]).

Conjecture. For each integer n it is true that

$$n \text{ is prime} \quad \Leftrightarrow \quad s_n := \sum_{k=1}^{n-1} k^{n-1} \equiv -1 \bmod n.$$

Fermat's little theorem says that if p is a prime, then $k^{p-1} \equiv 1 \bmod p$ for $k = 1, \dots, p-1$. Therefore, for each prime p, $s_p \equiv -1 \bmod p$. The question becomes:

Does there exist a non-prime n such that $s_n \equiv -1 \bmod n$?

This question has resisted solution for more than forty years. After surveying what is known about the conjecture, we will give several new results here which might suggest directions of further investigations.

The key to dealing with Giuga's conjecture is the following theorem, which was proved by Giuga in his original paper. A proof can also be found in [5]. For the sake of completeness, we give the proof here.

Theorem 1. *$s_n \equiv -1 \bmod n$ if and only if for each prime divisor p of n we have $(p-1)|(n/p-1)$ and $p|(n/p-1)$.*

Proof: It is well-known and an easy consequence of considering residue classes (see [5], p. 16) that for a prime p, we have

$$\sum_{k=1}^{p-1} k^{n-1} \equiv \begin{cases} -1 \bmod p & \text{if } (p-1)|(n-1), \\ 0 \bmod p & \text{if } (p-1) \nmid (n-1). \end{cases}$$

Therefore, for each prime divisor p of n with $n = p \cdot q$, we get

$$\sum_{k=1}^{n-1} k^{n-1} \equiv q \sum_{k=1}^{p-1} k^{n-1} \equiv \begin{cases} -q \bmod p & \text{if } (p-1)|(n-1), \\ 0 \bmod p & \text{if } (p-1) \nmid (n-1). \end{cases} \tag{1}$$

Assume that $s_n \equiv -1 \bmod n$. Then for each prime divisor p of n, $n = p \cdot q$, we have

$$-1 \equiv \begin{cases} -q \bmod p & \text{if } (p-1)|(n-1), \\ 0 \bmod p & \text{if } (p-1) \nmid (n-1). \end{cases} \tag{2}$$

This is only possible if $(p-1)|(n-1) = q(p-1) + (q-1)$. So $(p-1)|(q-1)$. It then also follows from (2) that $-1 \equiv -q \bmod p$, or $p|(q-1)$.

On the other hand, assume that $p|(q-1)$ and $(p-1)|(q-1)$. It then follows from (1) that $s_n \equiv -q \bmod p$; since $q \equiv 1 \bmod p$, we have that $s_n \equiv -1 \bmod p$ for each prime divisor p of n. Now, n must be squarefree: If it were not, then there would exist a prime divisor p of n with $p|q$; this contradicts $p|(q-1)$. Since

each of the distinct prime divisors of n divides $s_n + 1$, this is also true for n. In other words, $s_n \equiv -1 \bmod n$. □

As noted in the proof, every counterexample to Giuga's conjecture must be squarefree. Squarefree composite numbers which satisfy the first of these two conditions have been investigated in their own right: they are called *Carmichael numbers*. They were introduced by Carmichael in 1910. Carmichael numbers are of interest because they are "pseudo-prime" in the following sense (Korselt's criterion, 1899): *n divides $a^n - a$ for all integers a if and only if n is squarefree and $p - 1$ divides $n/p - 1$ for all prime divisors p of n.* The Carmichael condition

$$(p-1)|(n/p-1) \text{ for all prime divisors } p \text{ of } n$$

is equivalent to the condition

$$(p-1)|(n-1) \text{ for all prime divisors } p \text{ of } n.$$

Note that any Carmichael number is odd. (Assume that n is even. It has at least one other prime factor p besides 2. Then the even number $p - 1$ divides the odd number $n - 1$, which is a contradiction.) The smallest Carmichael number is $561 = 3 \cdot 11 \cdot 17$. The next two Carmichael numbers are $1105 = 5 \cdot 13 \cdot 17$ and $1729 = 7 \cdot 13 \cdot 19$. It has only recently been proved that there are infinitely many Carmichael numbers (see [**1**]).

In order to refer easily to the second condition as well, we will call any composite number n with $p|(n/p - 1)$ for all prime divisors p of n a *Giuga number*. As we saw in the proof of Theorem 1, any Giuga number is squarefree. Moreover, one can prove the following equivalence along the lines of that proof (with $n - 1$ replaced by $\varphi(n)$, where φ is the Euler (totient) function): *n is a Giuga number if and only if $\sum_{k=1}^{n-1} k^{\varphi(n)} \equiv -1 \bmod n$.* In his original paper, Giuga proved another equivalence: *n is a Giuga number if and only if*

$$\sum_{p|n} \frac{1}{p} - \prod_{p|n} \frac{1}{p} \in \mathbb{N}.$$

This equivalence will be of great importance throughout this paper. We will give a slightly generalized version of Giuga's proof below. The smallest Giuga number is 30: $1/2 + 1/3 + 1/5 - 1/30 = 1$. The next two Giuga numbers are 858: $1/2 + 1/3 + 1/11 + 1/13 - 1/858 = 1$, and 1722: $1/2 + 1/3 + 1/7 + 1/41 - 1/1722 = 1$. We do not know if there are infinitely many Giuga numbers.

Giuga's theorem can now be restated as

Theorem. *A composite integer n satisfies $s_n \equiv -1 \bmod n$ if and only if it is both a Carmichael number and a Giuga number.*

Giuga's conjecture is that such a number cannot exist.

The smallest odd Giuga number has at least 9 prime factors, since with a smaller number of prime factors the sum $1/p_1 + \ldots + 1/p_m - 1/n$ is smaller than 1. However, this lower bound for a counterexample increases dramatically if we take into account that it must also be a Carmichael number. Any Carmichael number n has the following property: If p is a prime factor of n, then for no k is $kp + 1$ a prime factor of n. (If it were, then we would have $(kp + 1) - 1 = kp|(n - 1)$ and $p|n$, which is a contradiction.) So, for example, no Carmichael number has the prime factors 3 and 7 at the same time. This property was used by Giuga to prove computationally that each counterexample has at least 1000 digits. Later, E. Bedocchi ([**2**]) used the same method to prove that each counterexample has at

least 1700 digits. We will describe the method in Section 2 of this article. We have been able to improve on this method by reducing the number of cases to be looked at and have shown computationally that any counterexample has no less than 13,800 digits.

We believe that an approach to prove or refute Giuga's conjecture in general is to study Giuga numbers in more depth; we will do this in Section 3 of this paper. It turns out that there is much more structure to be studied if we drop the condition that the numbers p in the definition be prime. This leads us to the following definition.

Definition. *A finite, increasing sequence of integers* $[n_1, \dots, n_m]$, *is called a Giuga sequence if*

$$\sum_{i=1}^{m} \frac{1}{n_i} - \prod_{i=1}^{m} \frac{1}{n_i} \in \mathbb{N}.$$

The proof of the following equivalence is due to Giuga in the case that the n_i are primes.

Theorem 2. *A finite, increasing sequence* $[n_1, \dots, n_m]$ *is a Giuga sequence if and only if it satisfies* $n_i | (n_1 \cdots n_{i-1} \cdot n_{i+1} \cdots n_m - 1)$ *for* $i = 1, \dots, m$.

Proof: Write $n := n_1 \cdots n_m$ and $q_i := n/n_i$. Note that the sequence is a Giuga sequence if and only if $n | (q_1 + \dots + q_m - 1)$.

The "only if" part of the asserted equivalence now follows immediately from this.

On the other hand, assume that $n_i | (q_i - 1)$ or $n_i^2 | (n - n_i)$ for all i. Multiplying leads to $n^2 | (n - n_1) \cdots (n - n_m)$. In evaluating this product we can drop all multiples of n^2. We therefore get $n^2 | (n(q_1 + \dots + q_m) - n_1 \cdots n_m) = n(q_1 + \dots + q_m - 1)$. Therefore, $n | (q_1 + \dots + q_m - 1)$, which means that $[n_1, \dots, n_m]$ is a Giuga sequence. □

Note that each two distinct elements n_i, n_j in a Giuga sequence are relatively prime ($n_i | (n/n_i - 1)$, but $n_j | n/n_i$). When n is a Giuga number, it gives rise to a Giuga sequence (its prime factors), but in general it is conceivable that an integer n can have two different factorizations, both of which are Giuga sequences. However, we know of no example of this. We do know that there is an infinity of Giuga sequences; we will show this in Section 3 of this article.

In the same way, Carmichael numbers can be generalized to Carmichael sequences: *A finite, increasing sequence,* $[a_1, \dots, a_m]$, *is called a Carmichael sequence if* $(a_i - 1) | (a_1 \cdots a_m - 1)$ *for* $i = 1, \dots, m$. Note that a Carmichael sequence has either exclusively odd or exclusively even elements and that its elements need not be relatively prime. As with Carmichael numbers, any two distinct factors a_i, a_j in a Carmichael sequence satisfy $a_i \not\equiv 1 \bmod a_j$ (or, equivalently, $a_j \nmid (a_i - 1)$). Carmichael sequences occur in much greater profusion than Giuga sequences. At the end of Section 3 we will construct some infinite families of Carmichael sequences.

Giuga's conjecture would be proved if one were to show that no Giuga sequence can be a Carmichael sequence. This, in turn, would be proved if it can be shown that any Giuga sequence must contain two factors n_i, n_j with $n_j | (n_i - 1)$. It might even be true that every Giuga sequence contains an even factor; Giuga's conjecture would follow from this. We have found no Giuga sequence which consists of odd

factors only, but this is probably a consequence of the size of the problem.

In Section 4 of this article, we will give a list of open questions concerning Giuga sequences and Giuga's conjecture.

2. COMPUTING LOWER BOUNDS FOR A COUNTEREXAMPLE. As we have seen, any counterexample to Giuga's conjecture must be a squarefree odd number with prime factorization $n = q_1 \cdots q_k$ such that

(i) $q_i \not\equiv 1 \bmod q_j$ for all i, j, and
(ii) $1/q_1 + \ldots + 1/q_k > 1$.

Giuga and Bedocchi used these two properties to compute lower bounds for a counterexample in the following way. For $m \in \mathbb{N}$, denote by p_m the m-th odd prime. A finite set of odd primes, $\{q_1, \ldots, q_k\}$, is called *normal* if condition (i) obtains. For each $m \in \mathbb{N}$, let S_m be the set of all normal sets with maximum element smaller than p_m. For each $S \in S_m$, $S = \{q_1, \ldots, q_k\}$, define the set

$$T_m(S) = \{q_1, \ldots, q_k, q_{k+1}, \ldots, q_r\}$$

to be the smallest set of odd primes which contains S, and is such that $q_j \geq p_m$ and $S \cup \{q_j\}$ is normal for $j > k$, and $\Sigma_{j=1}^{r} 1/q_j > 1$. Let $r_m(S)$ be the number of elements of $T_m(S)$. For example,

$$T_6(\{3,5\}) = \{3,5,17,23,29,47,53,\ldots,7919,7937\}, r_6(\{3,5\}) = 383$$

and

$$T_1(\{\ \}) = \{3,5,7,11,13,17,19,23,29\}, r_1(\{\ \}) = 9.$$

Now define the sequence $(i_m)_{m \in \mathbb{N}}$ by

$$i_m = \min\{r_m(S) | S \in S_m\}.$$

As Giuga observed, this sequence is non-decreasing; we will see shortly why this is the case.

Now, the number of prime factors of a counterexample, n, to Giuga's conjecture exceeds i_m for each $m \in \mathbb{N}$. Indeed, the prime factors form a normal set, and the subset S of the factors smaller than p_m is a member of S_m. Since any normal set of primes which contains S and satisfies condition (ii) above must contain at least $r_m(S)$ elements, we have that n has at least $r_m(S) \geq i_m$ prime factors; this is true for any $m \in \mathbb{N}$. So, any counterexample is bigger than $\Pi_{j=1}^{i_m} p_j$, and therefore has at least the same number of digits as this product. Giuga estimated $i_9 > 361$, this yields more than 1000 digits; Bedocchi computed $i_9 = 554$, this yields more than 1700 digits. (Note that Giuga and Bedocchi used a slightly different definition for i_m; this is why their numbers differ from our numbers by 1.)

To compute i_m, one has to find $r_m(S)$ for all $S \in S_m$. Since the number of elements of S_m increases geometrically with m, the time needed to compute i_m gets out of hand quickly. With our R4000 Challenge server and the symbolic manipulation package Maple, we were able, with considerable effort, to compute $i_{19} = 825$. At this point in time we started looking for a better algorithm, something which allows us to look at only some sets in S_m, not all of them. Fortunately enough, we found just such an algorithm. It is based on the following observation. Consider a set $S \in S_m$ and the associated value $r_m(S)$. Now, S has at most two "successors" in the set S_{m+1}, namely S itself and the set $S' = S \cup \{p_m\}$. We will now show that $r_{m+1}(S) \geq r_m(S)$ and $r_{m+1}(S') \geq r_m(S)$. In fact, there are two cases:

Case (i): $S \cup \{p_m\}$ is normal. Then S has the two successors S and S' in S_{m+1}. Also, we have $p_m \in T_m(S)$. However, $p_m \notin T_{m+1}(S)$, but every other element of $T_m(S)$ is contained in $T_{m+1}(S)$. So, $T_{m+1}(S)$ must contain at least one higher prime for the sum $\Sigma_{q \in T_{m+1}(S)} 1/q$ to exceed 1. Therefore, $r_{m+1}(S) \geq r_m(S)$. As regards S', the set $T_m(S)$ may contain primes which are congruent to $1 \bmod p_m$. These are missing in $T_{m+1}(S')$, since $p_m \in S'$. For each of these we need at least one higher prime for the sum $\Sigma_{q \in T_{m+1}(S')} 1/q$ to exceed 1. Again, $r_{m+1}(S') \geq r_m(S)$.

Case (ii): $S \cup \{p_m\}$ is not normal. Then the only successor of S in S_{m+1} is S itself. Also, $T_m(S) = T_{m+1}(S)$; the prime p_m is not contained in either set. Therefore, $r_m(S) = r_{m+1}(S)$.

This shows that the sequence i_m is indeed non-decreasing. But it shows more: the values $r_{k+1}, r_{k+2}, \ldots$ for all of the successors in $S_{k+1}, S_{k+2}, \ldots$ of a given set $S \in S_k$ do not fall below $r_k(S)$. If we want to compute i_m and already know an upper bound $I \geq i_m$, then we do not have to look at any successor in the sets $S_{k+1}, \ldots, S_m$ of a set $S \in S_k$ with $r_k(S) > I$. So the natural way to do this is to do it iteratively: Start with $A_1 := S_1$, and let A_{k+1} consist of the successors in S_{k+1} of all $S \in A_k$ with $r_k(S) \leq I$. Then $i_m = \min\{r_m(S) | S \in A_m\}$. If I is close to i_m, then this significantly reduces the number of sets to consider.

The bound I can of course be chosen as the value $r_m(S)$ for some $S \in S_m$. The iterative method saves the most time if one correctly guesses which sets have low values. By looking at preliminary computational results, we discovered that the following rule seems to hold. Let $L_5 := \{5, 7\}$, and define

$$L_{k+1} := \begin{cases} L_k \cup \{p_k\} & \text{if } L_k \cup \{p_k\} \text{ is normal,} \\ L_k & \text{otherwise.} \end{cases}$$

Then it seems that for $m \geq 5$, $r_m(L_m) = i_m$. We have no proof that this is always true, but having discovered that the sets L_m yield good upper bounds for i_m, we employed our iterative method with these upper bounds to compute all values i_m for $m \leq 100$ in Maple and later for $m \leq 135$ in C. We always found that $r_m(L_m) = i_m$.

It was, by the way, surprisingly difficult to translate what was a fairly straightforward Maple program into C. While Maple handled the data structures we required (lists of sets of variable length) easily, it was a nontrivial problem to implement these in C. We gained a speed-up of a factor of around 5 (for m around 100), though. Even with this speed-up, the last case ($m = 135$) took 303 cpu hours, and the "curse of exponentiality" makes further computation close to impracticable. We thank Gerald Kuch (now a graduate student at the University of Waterloo) for doing this conversion from Maple to C.

Here are some of the i_m (the first nine of these are also given by Bedocchi in [2]): $i_1 = 9$, $i_2 = 27$, $i_3 = 65$, $i_4 = 114$, $i_5 = 127$, $i_6 = 202$, $i_7 = 278$, $i_8 = 323$, $i_9 = i_{10} = i_{11} = 554$, $i_{12} = i_{13} = i_{14} = i_{15} = i_{16} = 704$, $i_{17} = i_{18} = 751$, $i_{19} = i_{20} = 825, \ldots, i_{49} = i_{50} = 2121, \ldots, i_{74} = i_{75} = 2657, \ldots, i_{99} = i_{100} = i_{101} = 3050, \ldots, i_{131} = i_{132} = i_{133} = i_{134} = i_{135} = 3459$. $i_{100} = 3050$ implies that any counterexample to Giuga's conjecture has at least 12055 digits, $i_{135} = 3459$ implies that any counter example has at least 13887 digits.

As Bedocchi points out in [2], this method is inherently incapable of showing that Giuga's conjecture holds for all integers: the set L_{27692} is normal, has 8135 elements and satisfies $\Sigma_{q \in L_{27692}} 1/q > 1$. Therefore, $i_m \leq 8135$ for all $m \geq 27692$.

3. GIUGA SEQUENCES. Recall that a Giuga sequence is a finite, increasing sequence of integers, $[n_1, \ldots, n_m]$, such that

$$\sum_{i=1}^{m} \frac{1}{n_i} - \prod_{i=1}^{m} \frac{1}{n_i} \in \mathbb{N}.$$

When dealing with Giuga's conjecture, we are mainly interested in Giuga sequences which consist exclusively of primes; we will call these *proper* Giuga sequences. However, Giuga sequences, proper or not, are interesting objects in their own right; in this section we will give some of their properties.

We have computed all Giuga sequences up to length 7 and some of length 8. There is no Giuga sequence of length 2; one sequence of length 3 ($[2, 3, 5]$); two sequences of length 4 ($[2, 3, 7, 41]$ and $[2, 3, 11, 13]$); three sequences of length 5 ($[2, 3, 7, 43, 1805]$, $[2, 3, 7, 83, 85]$ and $[2, 3, 11, 17, 59]$); 17 sequences of length 6; 27 sequences of length 7; and hundreds of sequences of length 8.

So far, we know only 11 Giuga numbers (or proper Giuga sequences). They are

3 factors:

$$30 = 2 \cdot 3 \cdot 5$$

4 factors:

$$858 = 2 \cdot 3 \cdot 11 \cdot 13$$
$$1722 = 2 \cdot 3 \cdot 7 \cdot 41$$

5 factors:

$$66198 = 2 \cdot 3 \cdot 11 \cdot 17 \cdot 59$$

6 factors:

$$2214408306 = 2 \cdot 3 \cdot 11 \cdot 23 \cdot 31 \cdot 47057$$
$$24423128562 = 2 \cdot 3 \cdot 7 \cdot 43 \cdot 3041 \cdot 4447$$

7 factors:

$$432749205173838 = 2 \cdot 3 \cdot 7 \cdot 59 \cdot 163 \cdot 1381 \cdot 775807$$
$$14737133470010574 = 2 \cdot 3 \cdot 7 \cdot 71 \cdot 103 \cdot 67213 \cdot 713863$$
$$550843391309130318 = 2 \cdot 3 \cdot 7 \cdot 71 \cdot 103 \cdot 61559 \cdot 29133437$$

8 factors:

$$244197000982499715087866346$$
$$= 2 \cdot 3 \cdot 11 \cdot 23 \cdot 31 \cdot 47137 \cdot 28282147 \cdot 3892535183$$
$$554079914617070801288578559178$$
$$= 2 \cdot 3 \cdot 11 \cdot 23 \cdot 31 \cdot 47059 \cdot 2259696349 \cdot 110725121051$$

For all of these examples, the 'sum minus product' value is 1; to reach any higher value, the sequence would have to have at least 59 factors. To find all Giuga sequences of a given length, one could check all sequences of this length whose elements are not too large (the sum over their reciprocals must be greater than 1). However, the number of these grows exponentially; even for length 7 there are too many to check them all. Fortunately, we have the following theorem which tells us how to find all Giuga sequences of length m with a given initial segment of length $m - 2$.

Theorem 3. (*a*) *Take an initial sequence of length* $m - 2, [n_1, \ldots, n_{m-2}]$. *Let*

$$P = n_1 \cdots n_{m-2}, S = 1/n_1 + \ldots + 1/n_{m-2}.$$

Fix an integer $v > S$ (*this will be the sum minus product value*). *Take any integers* a, b

with $a \cdot b = P(P + S - v)$ *and* $b > a$. *Let*

$$n_{m-1} := (P + a)/P(v - S), n_m := (P + b)/P(v - S).$$

Then

$$S + 1/n_{m-1} + 1/n_m - 1/P_{n_{m-1}} n_m = v.$$

Then sequence $[n_1, \dots, n_{m-1}, n_m]$ *is a Giuga sequence if and only if* n_{m-1} *is an integer.*

(*b*) *Conversely, if* $[n_1, \dots, n_{m-1}, n_m]$ *is a Giuga sequence with sum minus product value* v, *and if we define*

$$a := n_{m-1}P(v - S) - P, b := n_m P(v - S) - P$$

(*with* P *and* S *the product and the sum of the first* $m - 2$ *terms*) *then* a *and* b *are integers and* $a \cdot b = P(P + S - v)$.

Proof: (a) First we have to check that with these definitions for n_{m-1} and n_m we have in fact $S + 1/n_{m-1} + 1/n_m - 1/Pn_{m-1}n_m = v$. These are straight-forward calculations:

$$S + P(v - S)/(P + a) + P(v - S)/(P + b) \\ - P^2(v - S)^2/P(P + a)(P + b) = v$$

if and only if

$$P/(P + a) + P/(P + b) - P(v - S)/(P + a)(P + b) = 1$$

if and only if

$$P/(P + a) + P/(P + b) - (P^2 - ab)/(P + a)(P + b) = 1,$$

which is true.

This means that the completed sequence is a Giuga sequence if and only if both n_{m-1} and n_m are integers. It remains to be shown that n_{m-1} is integer if and only if n_m is. Because of the symmetry, it is enough to prove the implication in one direction. If

$$P(v - S) | (P + a)$$

then

$$P(v - S) | (P + a)(P + b) = 2P^2 + (a + b)P - (v - S)P,$$

so

$$P(v - S) | (2P^2 + (a + b)P) = P(P + a + P + b),$$

so

$$P(v - S) | P(P + b)$$

(since $P(v - S)$ divides $P + a$).

The assertion follows if we show that $\gcd(P(v - S), P) = 1$. Assume that there is a prime p with $p | P(v - S)$ and $p | P$, $p | n_i$, say. Since $\gcd(n_i, n_j) = 1$ for i not equal to j, p does not divide any of the other factors. Since p divides

$$P(v - S) = vP - (n_2 \cdots n_{m-2} + \dots + n_1 \cdots n_{m-3}),$$

we can drop all terms on the right-hand side with a factor n_i to get

$$p | n_1 \cdots n_{i-1} \cdot n_{i+1} \cdots n_{m-2},$$

which is a contradiction.

(b) It is clear that these a and b are integers. It remains to be checked that $a \cdot b = P(P + S - v)$. We have

$$\begin{aligned} a \cdot b &= -P^2(v - S)(n_{m-1} + n_m) + P^2 n_{m-1} n_m (v - S)^2 + P^2 \\ &= P^2(S - v) n_{m-1} n_m \cdot (1/n_{m-1} + 1/n_m + S - v) + P^2 \\ &= P^2(S - v) n_{m-1} n_m \cdot 1/(n_{m-1} n_m P) + P^2 \\ &= P(P + S - v). \end{aligned}$$

□

We used this theorem to compute our Giuga sequences; this works well for sequences up to length 7, but for length 8 and higher there are still too many possible initial segments to check.

All Giuga sequences we have found so far contain an even factor (it is usually the factor 2, but there are two sequences of length 9 which contain 4 instead). So far, we have not been able to find a sequence with only odd factors. Since any counterexample to Giuga's conjecture would be an odd Giuga number, it would be of some interest to find at least one such sequence. Let $n := n_1 \cdots n_m$. The Giuga equation $1/n_1 + \ldots + 1/n_m - 1/n = v$ is equivalent to $n/n_1 + \ldots n/n_m - 1 = nv$; by considering this equation modulo 4, it is quite straightforward to show that if all factors n_i are odd, then necessarily $m - v \equiv 1 \bmod 4$. (In fact, assume that the first k factors are congruent to -1 modulo 4, and the other $m - k$ factors are congruent to 1 modulo 4. Then the equation reduces to

$$-1 \equiv v(-1)^k - k(-1)^{k-1} - (m - k)(-1)^k = (-1)^k(v - m + 2k) \bmod 4,$$

from which $m - v \equiv 1 \bmod 4$ follows.) If we look for odd sequences with value $v = 1$, then we only have to check sequences of length $m \equiv 2 \bmod 4$. The cases $m = 2$ and $m = 6$ can be ruled out because we need at least nine relatively prime odd integers for the sum of their reciprocals to exceed 1. Now, $m = 10$ can be ruled out computationally, with the use of Theorem 3. But this is where computational feasibility ends. For $m = 14$, there are just too many initial segments to check; another approach is needed here.

We asserted in the introduction that there are infinitely many Giuga sequences. As the following proposition tells us, it is possible to generate longer Guiga sequences out of shorter ones with certain properties.

Theorem 4. *Take a Giuga sequence of length m, $[n_1, \ldots, n_{m-1}, n_m]$, which satisfies*

$$n_m = n_1 \cdots n_{m-1} - 1. \tag{3}$$

Let

$$\tilde{n}_m := n_1 \cdots n_{m-1} + 1, \tilde{n}_{m+1} := n_1 \cdots n_{m-1} \tilde{n}_m - 1.$$

Then $[n_1, \ldots, n_{m-1}, \tilde{n}_m, \tilde{n}_{m+1}]$ is also a Giuga sequence with the same sum minus product value.

Proof: Let $P := n_1 \cdots n_{m-1}$, $S := 1/n_1 + \ldots + 1/n_{m-1}$. Then $n_m = P - 1$, $\tilde{n}_m = P + 1$ and $\tilde{n}_{m+1} = P^2 + P - 1$. Both sequences have the same sum minus product value if and only if

$$S + \frac{1}{P - 1} - \frac{1}{P(P - 1)} = S + \frac{1}{P + 1} + \frac{1}{P^2 + P - 1} - \frac{1}{P(P + 1)(P^2 + P - 1)};$$

the latter equation is true for all S and P. □

Note that if the shorter Giuga sequence has property (3), then so has the longer one. Since the sequence [2, 3, 5] also has this property, this proves that there are Giuga sequences of any length. (Note also that each of the sequences which occur in such a recursion will contain an even factor.) However, the sequences arrived at by this recursion are not the only Giuga sequences there are. In each step from length m to $m + 1$, other Giuga sequences seem to pop out of thin air, some of them with property (3) (and thus leading to new recursions), some without.

Explicitly, two infinite families are

(a) $n_1 = 2, n_k = n_1 \cdots n_{k-1} + 1$ for $k = 2,, \ldots, m - 1, n_m = n_1 \cdots n_{m-1} - 1$;
(b) $n_1 = 2, n_2 = 3, n_3 = 11, n_4 = 23, n_5 = 31, n_k = n_1 \cdots n_{k-1} + 1$ for $k = 6, \ldots, m - 1, n_m = n_1 \cdots n_{m-1} - 1$.

Recall that a Carmichael sequence is a finite increasing sequence of integers, $[a_1, \ldots, a_m]$, such that

$$(a_i - 1) \Bigg| \left(\prod_{j=1}^{m} a_j - 1 \right) \text{ for } i = 1, \ldots, m.$$

At the end of the introduction we stated that there is a profusion of Carmichael sequences; in fact, there are infinitely many of them with any number of factors. A trivial example would be the sequence $[a, \ldots, a]$ for any $a \in \mathbb{N}$, but this is too cheap. We now conclude this section by giving a somewhat less trivial construction. We omit the computations here, since they are essentially simple but would enlarge this side remark unduly.

There are infinitely many Carmichael sequences of length 3; in fact, the following construction gives us *all* 3-factor Carmichael sequences. Take three integers $b_1, b_2, b_3 \in \mathbb{N}$ which are pairwise co-prime. Let $c \in \mathbb{N}$ be the solution of the congruence $(b_1 b_2 + b_1 b_3 + b_2 b_3)c \equiv -(b_1 + b_2 + b_3) \bmod b_1 b_2 b_3$. (Such a solution exists and is unique modulo $b_1 b_2 b_3$, because the integers are pairwise co-prime. Equivalently, one can also solve the system of congruences $b_2 b_3 c \equiv -b_2 - b_3 \bmod b_1, b_1 b_3 c \equiv -b_1 - b_3 \bmod b_2, b_1 b_2 c \equiv -b_1 - b_2 \bmod b_3$). Then $a_1 := cb_1 + 1, a_2 := cb_2 + 1, a_3 := cb_3 + 1$ is always a Carmichael sequence, and every 3-factor Carmichael sequence is of this form. For example, if we choose $b_1 = 1$, $b_2 = 2$ and $b_3 = 3$, then we get Chernick's observation (see [1]) that $a_1 := 6k + 1, a_2 := 12k + 1, a_3 := 18k + 1$ is always a Carmichael sequence.

The following two recursions produce Carmichael sequences of length $m + 1$ and $m + 2$ out of the Carmichael sequence $[a_1, \ldots, a_m]$.

(a) Let $a_{m+1} := \prod_{j=1}^{m} a_j$. Then $[a_1, \ldots, a_m, a_{m+1}]$ is a Carmichael sequence.
(b) Let $a_{m+1} := \prod_{j=1}^{m} a_j$ and $a_{m+2} := d(a_{m+1} - 1) + 1$ where d is a divisor of $a_{m+1} + 1$. Then $[a_1, \ldots, a_m, a_{m+1}, a_{m+2}]$ is a Carmichael sequence.

In particular, if we start with an odd 3-factor Carmichael sequence, then we can extend it to arbitrary length by iterating the step (b) with $d = 2$.

4. OPEN PROBLEMS

1. Giuga's conjecture: *Show that no integer exists which is both a Giuga number and a Carmichael number*. More general: *Show that no Giuga sequence can be a Carmichael sequence*. (This would imply the truth of Giuga's conjecture.)

2. Does every Giuga sequence contain two factors n_i, n_j with $n_j|(n_i - 1)$? If this were true, then Giuga's conjecture would be proved.
3. Find a Giuga sequence which consists of odd factors (or odd primes) only, or prove that none exist. If there were none, then Giuga's conjecture would be proved.
4. Are there infinitely many proper Giuga sequences?
5. Is it true that, in the notation of Section 2, $r_m(L_m) = i_m$ for $m = 5, \dots, 27692$? If so, then this would prove that a counterexample to Giuga's conjecture has at least 36069 digits.
6. Find a fast way to generate all Giuga sequences of a given length.
7. Are there Giuga sequences with a sum minus product value higher than 1?
8. Are there two distinct Giuga sequences whose elements have the same product?
9. Can each integer be a factor in a Giuga sequence? If this were true then it would answer the previous two questions positively. In fact, take any integer n which is the product of a Giuga sequence, $n = n_1 \cdots n_m$ with $1/n_1 + \dots + 1/n_m - 1/n = v$. If we can find a second Giuga sequence which contains n as a factor, e.g., $1/n + 1/\tilde{n}_1 + \dots + 1/\tilde{n}_k - 1/n\tilde{n}_1 \cdots \tilde{n}_k = w$, then we can combine the two of them and get the Giuga sequence $1/n_1 + \dots + 1/n_m + 1/\tilde{n}_1 + \dots 1/\tilde{n}_k - 1/n_1 \cdots n_m\tilde{n}_1 \cdots \tilde{n}_k = v + w$. It has the same product as the previous one, but a higher sum minus product value.
10. Agoh's conjecture: *Let B_k denote the kth Bernoulli number. Then $nB_{n-1} \equiv -1 \bmod n$ if and only if n is a prime*? Note that the denominator of the number nB_{n-1} can be greater than 1, but since the denominator of any Bernoulli number is squarefree, the denominator of nB_{n-1} is invertible modulo n. As Takashi Agoh (Science University of Tokyo) has informed us, this recent conjecture of his is equivalent to Giuga's conjecture: Every counterexample of Giuga's conjecture is also a counterexample to Agoh's conjecture and vice versa. This can be seen from the well-known formula

$$s_{n-1} = \sum_{i=1}^{n} \binom{n}{i} n^{i-1} B_{n-i}$$

after some analysis involving von Staudt-Clausen's theorem: *The denominator of B_{2k} is given by* $\prod_{\substack{p \text{ prime} \\ (p-1)|2k}} p$. (See [4], pp. 91–93; this also implies that the denominator of B_{2k} is squarefree.) Incidentally, it is possible to use a similar argument to characterize Giuga numbers in the following way: *n is a Giuga number if and only if $nB_{\varphi(n)} \equiv -1 \bmod n$.*

Finally, we would like to thank Hugh Edgar for originally making us aware of Giuga's conjecture and challenging us to extend what was known computationally.

REFERENCES

1. W. R. Alford, A. Granville, C. Pomerance, There are infinitely many Carmichael numbers, *Ann. of Math.* **140** (1994), 1–20.
2. E. Bedocchi, Nota ad una congettura sui numeri primi, *Riv. Mat. Univ. Parma* **11** (1985), 229–236.
3. G. Giuga, Su una presumibile proprietà caratteristica dei numeri primi, *Ist. Lombardo Sci. Lett. Rend. A* **83** (1950), 511–528.

4. G. H. Hardy, E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford 1959.
5. P. Ribenboim, *The Book of Prime Number Records*, Springer-Verlag 1989.

D. Borwein:
Department of Mathematics
University of Western Ontario
London, Ontario
N6A 5B7, Canada
dborwein@uwo.ca

J. M. Borwein:
Department of Mathematics and Statistics
Simon Fraser University
Burnaby BC V5A 1S6, Canada
jborwein@cecm.sfu.ca

P. Borwein:
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
pborwein@cecm.sfu.ca

R. Girgensohn:
Institut für Mathematik
Medizinische Universität Lübeck
P-23560 Lübeck
GERMANY
girgenso@informatik.mu-luebeck.de

Rounding Error?

From article #155 about rocks in Antarctica in the book *365 Surprising Scientific Facts, Breakthroughs, and Discoveries* by S. B. McGrayne, Wiley, 1994:

...north-facing rocks that soak up polar sunlight can be 15°C (60°F) warmer than the surrounding air.

Submitted by H. Turner Laquer
Idaho State University

Answer to Picture Puzzle

(p. 29)

Jack Schwartz.

A Geometric Interpretation of the Solution of the General Quartic Polynomial

William M. Faucette

Suppose we are given the general polynomial equation of degree n:

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0,$$

where each of the a_i's is a rational number and a_n is not zero. We might ask if the solutions of this equation can be expressed in terms of the coefficients $a_0, \dots, a_n$ using only the operations of addition, subtraction, multiplication, division, and extraction of roots. One of the principal results of Galois Theory, Abel's theorem, states that such formulas exist for $n \leq 4$ and do not exist for $n \geq 5$. The reader can find a discussion of Abel's theorem in numerous sources, including **[A]**, **[F1]**, **[H1]**, and **[H2]**.

In this article we will first recall the explicit radical solution of cubic polynomials. We will then proceed to discuss the solution of the general quartic polynomial by reduction to an auxiliary cubic equation, the quartic's resolvent cubic. The algebraic solutions presented here appear in section 4.16 of the text **[E]**.

After defining algebraic plane curves and introducing a few facts about them, we will present an interesting algebro-geometric interpretation of the derivation of the resolvent cubic.

I would like to express thanks to Miles Reid for his text [**R**, pp. 22–24] which inspired this article. I would also like to thank Elsa Newman, Bill Rivera, David Smead, Chris Vaughn, and the referee, whose suggestions contributed to the readability of the finished product.

GALOIS SOLUTION OF THE GENERAL CUBIC POLYNOMIAL. Let $P(z) = z^3 + a_1 z^2 + a_2 z + a_3$ be a cubic polynomial with rational coefficients. To simplify the solution we eliminate the quadratic term by setting $z = x - \frac{1}{3}a_1$. Then $P(z)$ takes the form $\tilde{P}(x) = x^3 + px + q$, where p and q are polynomials in the coefficients of $P(z)$. Notice that solving $\tilde{P}(x)$ readily solves $P(z)$.

Let x_1, x_2, and x_3 be the roots of $\tilde{P}(x)$, which we assume to be distinct. Notice that since $\tilde{P}(x) = (x - x_1)(x - x_2)(x - x_3)$ has no quadratic term, the sum of the roots must be zero. Let ω be a primitive cube root of unity and define the *Lagrange resolvents*, $(1, x_1)$, (ω, x_1), and (ω^2, x_1), by

$$\begin{aligned}(1, x_1) &= x_1 + x_2 + x_3 = 0\\ (\omega, x_1) &= x_1 + \omega x_2 + \omega^2 x_3 \qquad (1)\\ (\omega^2, x_1) &= x_1 + \omega^2 x_2 + \omega x_3.\end{aligned}$$

Algebraic manipulation shows that the Lagrange resolvents can be computed in terms of the coefficients of $\tilde{P}(x)$ and the square root of the discriminant of $\tilde{P}(x)$. Solving equations (1) for x_1, x_2, and x_3 gives the roots of $\tilde{P}(x)$ in terms of the Lagrange resolvents. Substituting the value of the Lagrange resolvents into the solutions of (1) yields the zeroes of $\tilde{P}(x)$, from which the zeroes of $P(z)$ can be obtained.

GALOIS SOLUTION OF THE GENERAL QUARTIC POLYNOMIAL. Consider the general quartic with rational coefficients, given by $P(z) = z^4 + a_1 z^3 + a_2 z^2 + a_3 z + a_4$. As with the cubic, we first simplify the polynomial by the substitution $z = x - \frac{1}{4}a_1$, yielding

$$\tilde{P}(x) = x^4 + px^2 + qx + r \tag{2}$$

where p, q, and r are polynomials in the coefficients of $P(z)$.

Let x_1, x_2, x_3, and x_4 be the roots of $\tilde{P}(x)$. Since $\tilde{P}(x) = (x - x_1)(x - x_2)(x - x_3)(x - x_4)$ has no cubic term, the sum of the roots once again must be zero. Define $\alpha = (x_1 + x_2)(x_3 + x_4)$, $\beta = (x_1 + x_3)(x_2 + x_4)$, $\gamma = (x_1 + x_4)(x_2 + x_3)$. Let $h \in \mathbb{Q}[z]$ be the polynomial $h = (z - \alpha)(z - \beta)(z - \gamma)$, the *resolvent cubic of* $\tilde{P}(x)$. A little calculation shows that $h = z^3 - 2pz^2 + (p^2 - 4r)z + q^2$.

By solving this cubic equation using the method in the preceding section, one obtains α, β, and γ. Using

$$\begin{aligned}
0 &= (x_1 + x_2) + (x_3 + x_4) \text{ and } \alpha = (x_1 + x_2)(x_3 + x_4)\\
0 &= (x_1 + x_3) + (x_2 + x_4) \text{ and } \beta = (x_1 + x_3)(x_2 + x_4)\\
0 &= (x_1 + x_4) + (x_2 + x_3) \text{ and } \gamma = (x_1 + x_4)(x_2 + x_3),
\end{aligned}$$

one obtains roots of $\tilde{P}(x)$. The zeroes of the original quartic may then be easily obtained. For complete algebraic solutions of the general cubic and quartic polynomials, see [**E**, §4.16], [**W**, §64], and [**B**, 16.4.10 and 16.4.11.1].

EVERYTHING YOU NEED TO KNOW ABOUT ALGEBRAIC PLANE CURVES. To give an algebro-geometric interpretation of the resolvent cubic, we need to introduce a few basic facts about algebraic curves. For a complete introduction to algebraic plane curves, see the text [**F2**].

Let $\mathbb{C}$ denote the field of complex numbers and define the affine complex plane, $\mathbb{A}^2$, to be the set of all ordered pairs (a, b) where $a, b \in \mathbb{C}$. A complex affine plane curve is the locus of zeroes in $\mathbb{A}^2$ of a nonzero polynomial $f \in \mathbb{C}[X, Y]$. The complex projective plane, $\mathbb{P}^2$, is the set of all equivalence classes $[a, b, c]$ of ordered triples $(a, b, c) \in \mathbb{C}^3 \setminus (0, 0, 0)$ under the equivalence relation $(a, b, c) \sim (a', b', c')$ if $(a, b, c) = (\lambda a', \lambda b', \lambda c')$ for some nonzero complex number λ. Notice that if $c \neq 0$, we may divide the three coordinates by c and obtain coordinates $[a, b, 1]$. A complex projective plane curve is the locus of zeroes in $\mathbb{P}^2$ of a nonzero homogeneous polynomial $F \in \mathbb{C}[X, Y, Z]$. The degree of a plane curve is the degree of its defining polynomial. Curves of degrees one, two, three, and four are called lines, conics, cubics, and quartics, respectively.

The affine plane is contained in the projective plane by the inclusion $\mathbb{A}^2 \hookrightarrow \mathbb{P}^2$ given by $(x, y) \mapsto [x, y, 1]$, with the remainder of the projective plane forming the line at infinity, $L_\infty = \{[x, y, 0] \in \mathbb{P}^2\}$. If $f(X, Y)$ is an element of $\mathbb{C}[X, Y]$ of degree d, we can homogenize f by setting $F(X, Y, Z) = Z^d f(X/Z, Y/Z)$. F is then a homogeneous polynomial of degree d. If f defines an affine plane curve C, the projective plane curve defined by F is the *projective closure* of C.

A general conic in $\mathbb{P}^2$ is given as the set of zeroes of an equation

$$F(X,Y,Z) = aX^2 + bXY + cY^2 + dXZ + eYZ + fZ^2, \tag{3}$$

where at least one of these coefficients is nonzero, and this equation is unique up to multiplication by a nonzero constant. A conic with equation (3) is reducible if and only if the equations

$$F_X(X,Y,Z) = F_Y(X,Y,Z) = F_Z(X,Y,Z) = 0$$

have a common solution in $\mathbb{P}^2$, where we use the subscripts to denote partial derivatives. If we let $\mathbf{A}$ be the matrix associated with (3), then

$$\mathbf{A} = \begin{bmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{bmatrix},$$

and we can rewrite equation (3) as

$$[X \quad Y \quad Z]\mathbf{A}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = 0.$$

The condition that the conic be reducible is equivalent to the condition that this associated matrix $\mathbf{A}$ is singular.

The set of all conics in $\mathbb{P}^2$ forms a five-dimensional projective space $\mathbb{P}^5$ in the following way. A general conic in $\mathbb{P}^2$ is given by an equation of the form (3), where at least one of these coefficients is nonzero, and this equation is unique up to multiplication by a nonzero constant. So, we may identify this conic with the point $[a, b, c, d, e, f] \in \mathbb{P}^5$. From this perspective, the conics in $\mathbb{P}^2$ passing through a given point P in $\mathbb{P}^2$ form a codimension one linear subspace in $\mathbb{P}^5$. That is, if $P = [u, v, w]$, then any conic through P must satisfy $F(u, v, w) = au^2 + buv + cv^2 + duw + evw + fw^2 = 0$, and this is a linear equation in a, b, c, d, e, f. Similarly the condition for a conic to contain points $P_1, P_2, P_3, P_4 \in \mathbb{P}^2$ is given by a system of four linear equations in a, b, c, d, e, f. From elementary linear algebra, the family of conics containing all four points will be a one-dimensional linear subspace of $\mathbb{P}^5$ exactly when these four conditions are linearly independent. We then have the following proposition:

Proposition. *The family of conics containing the distinct points P_1, P_2, P_3, and P_4 is (projective) one-dimensional if and only if P_1, P_2, P_3, and P_4 are noncollinear.*

Proof: Suppose the points are noncollinear. Without loss of generality, we may assume that P_1, P_2, and P_3 are noncollinear. It is sufficient to show there exists a conic containing $P_1, \dots, P_t$ and not containing $P_{t+1}, \dots, P_4$ for $t = 1, 2, 3$.

To produce a conic through P_1 and not through P_2, P_3, and P_4, choose two lines through P_1 and not containing P_2, P_3, or P_4. The union of these two lines is a reducible conic containing P_1 and not containing P_2, P_3, or P_4.

Let l be any line through P_1 not containing P_3 or P_4. Let l' be any line through P_2 not containing P_3 or P_4. The union of l and l' is a reducible conic containing P_1 and P_2, but not containing P_3 and P_4.

We divide the last part of the proof into two cases depending on the relative positions of P_1, P_2, and P_4. First, suppose P_1, P_2, and P_4 are noncollinear. Choose any line l' through P_3 not containing P_4. The union of l' and the line through P_1 and P_2 is then a reducible conic containing P_1, P_2, and P_3 and not P_4. On the

other hand, suppose P_1, P_2, and P_4 are collinear. Let l be the line through P_1 and P_3 and let l' be the line through P_2 and P_3. Then the union of l and l' is a reducible conic containing P_1, P_2, and P_3, and not P_4. This shows the family of conics containing P_1, P_2, P_3, and P_4 has dimension one. In this context, a linear subspace of dimension one is called a *pencil*, so this family is a pencil of conics.

Conversely, if P_1, P_2, P_3, and P_4 are collinear, let l be the line containing these four points. Let l' be any line in $\mathbb{P}^2$. Then the union of l and l' is a reducible conic containing all four points. Since l' is an arbitrary line in $\mathbb{P}^2$, this family has dimension two. □

Now we wish to investigate briefly the number of points of intersection of two projective plane curves of various degrees. First, if we intersect a projective line with a conic, we always get two points if the points are counted properly. To see this, we can parametrize any line in the projective plane by

$$\begin{aligned} X &= a_1 s + b_1 t \\ Y &= a_2 s + b_2 t \\ Z &= a_3 s + b_3 t, \end{aligned} \tag{4}$$

where s and t cannot both be zero. Substituting these equations into the equation of a general conic gives a homogeneous quadratic polynomial in s and t. Setting this polynomial equal to zero and solving yields two points $[s, t]$ in the projective line $\mathbb{P}^1$. Substituting back into equations (4) yields the two points where the line meets the conic.

If we similarly investigate the intersection of two conics in the projective plane, we find that two conics always meet in four points if the points are counted properly. If one of the conics is reducible, this result follows from the previous paragraph, so we may assume the conics are nonsingular. Choose coordinates in the projective plane so that one conic has projective equation $XZ = Y^2$. We then parametrize this conic by the equations

$$\begin{aligned} X &= s^2 \\ Y &= st \\ Z &= t^2, \end{aligned} \tag{5}$$

where once again s and t cannot both be zero. Substituting these equations into the equation of a general conic gives a homogeneous quartic polynomial in s and t. Setting this polynomial equal to zero and solving yields four points $[s, t]$ in the projective line $\mathbb{P}^1$. Substituting back into equations (5) yields the four points where the two conics meet.

These two elementary computations are special cases of a more general result known as Bézout's Theorem, which says that projective algebraic curves of degrees m and n having no common component always meet in mn points if the points are counted properly. For our purposes, the two cases outlined above suffice.

A GEOMETRIC SOLUTION TO THE GENERAL QUARTIC. Let's go back to the reduced quartic polynomial given in equation (2):

$$x^4 + px^2 + qx + r = 0,$$

where $p, q, r \in \mathbb{Q}$. Considering these polynomials as having complex coefficients and setting $y = x^2$, we see that the solutions to equation (2) are the x-coordinates of the points of intersection of the conics with affine equations

$$\begin{aligned} y^2 + py + qx + r &= 0 \\ y - x^2 &= 0, \end{aligned}$$

in the affine plane $\mathbb{A}^2$. If we take the projective closure of these curves in $\mathbb{P}^2$, we get the projective curves C_1 and C_2 defined by polynomials

$$F_1(x, y, z) = y^2 + pyz + qxz + rz^2$$

$$F_2(x, y, z) = yz - x^2,$$

respectively. Using Bézout's Theorem, the curves C_1 and C_2 meet in the four points P_1, P_2, P_3, P_4, all of which lie in the finite plane and have affine coordinates $P_i = (x_i, x_i^2)$.

To see that the conditions imposed by P_1, P_2, P_3, P_4 are independent, we need only show that these points are noncollinear in $\mathbb{P}^2$. However, the four distinct points P_1, P_2, P_3, P_4 all lie on the irreducible conic $y = x^2$ in the affine plane, so they are not collinear, again by Bézout's theorem. It follows from the proposition that the set of conics in $\mathbb{P}^2$ containing P_1, P_2, P_3, P_4 forms a (projective) one-dimensional linear subspace Π of $\mathbb{P}^5$, so the conics C_1 and C_2 span Π. That is, any curve C in Π has equation $\lambda F_1 + \mu F_2 = 0$, where either λ or μ is not zero.

We now wish to find those conics C in the linear family Π that are reducible.

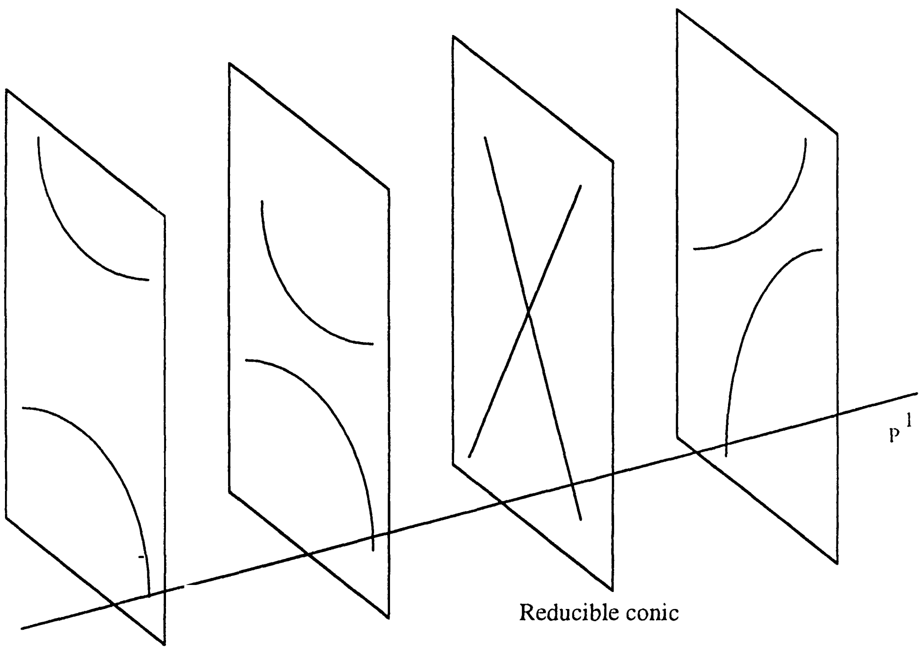


Figure 1. A Pencil of Conics Showing a Reducible Conic

The matrices $\mathbf{A}_i$ of conics C_i are given by

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & q/2 \\ 0 & 1 & p/2 \\ q/2 & p/2 & r \end{bmatrix} \qquad \mathbf{A}_2 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix},$$

so the matrix of the polynomial $F = \lambda F_1 + \mu F_2$ of any conic C in Π is given by

the matrix

$$\begin{bmatrix} -\mu & 0 & \frac{1}{2}q\lambda \\ 0 & \lambda & \frac{1}{2}p\lambda + \frac{1}{2}\mu \\ \frac{1}{2}q\lambda & \frac{1}{2}p\lambda + \frac{1}{2}\mu & r\lambda \end{bmatrix},$$

and C is reducible precisely when this matrix is singular. The determinant of this matrix is

$$\tfrac{1}{4}\left[\mu^3 - q^2\lambda^3 + (p^2 - 4r)\lambda^2\mu + 2p\lambda\mu^2\right]. \tag{6}$$

As the reader can see, this equation is homogeneous in λ and μ of degree three, so the roots $[\lambda, \mu]$ of this equation correspond to three reducible conics in the family Π. Let L_{ij} be the line through P_i and P_j. Then L_{ij} has affine equation $Y = (x_i + x_j)X - x_i x_j$. One of the three reducible conics in the family Π is $L_{12} + L_{34}$, which satisfies the polynomial

$$\begin{aligned}
&[Y - (x_1 + x_2)X + x_1x_2][Y - (x_3 + x_4)X + x_3x_4] \\
&\quad = Y^2 + (x_1x_2 + x_3x_4)Y + (x_1 + x_2)(x_3 + x_4)X^2 + qX + r \\
&\quad = F_1 - (x_1 + x_2)(x_3 + x_4)F_2,
\end{aligned}$$

noting that, by assumption, the coefficient of the XY term is $-(x_1 + x_2 + x_3 + x_4) = 0$. Hence, one of the roots of polynomial (6) is $[1, -(x_1 + x_2)(x_3 + x_4)] = [1, -\alpha]$. Similarly, the remaining two roots of polynomial (6) are $[1, -\beta]$ and $[1, -\gamma]$, so that the solutions of the resolvent cubic correspond geometrically to finding the three reducible conics in the space of conics spanned by C_1 and C_2.

Since the reducible conics in Π are

$$\begin{aligned}
Q_1 &= L_{12} + L_{34} \\
Q_2 &= L_{13} + L_{24} \\
Q_3 &= L_{14} + L_{23},
\end{aligned}$$

it is easy to see that the intersection of any two of these conics produces the desired points P_1, P_2, P_3, P_4.

Thus, if we interpret the roots of the general quartic as the first coordinates of points P_1, P_2, P_3, P_4 in the intersection of two conics in $\mathbb{P}^2$, we see that the resolvent cubic obtained from Galois Theory is, up to a nonzero constant multiple, just the determinant of the 3×3 matrix defining any conic in the family of conics containing the four points P_1, P_2, P_3, P_4. Solving the resolvent cubic corresponds geometrically to finding the reducible conics in this family. It is then a straightforward matter to solve the quartic equation geometrically.

REFERENCES

[A] E. Artin, *Galois Theory*, University of Notre Dame Press, Notre Dame, 1944.

[B] M. Berger, *Geometry*, Springer-Verlag, New York, 1987.

[E] G. Ehrlich, *Fundamental Concepts of Abstract Algebra*, PWS-Kent Publishing Company, Boston, 1991.

[F1] J. B. Fraleigh, *A First Course in Abstract Algebra*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1982.

[F2] W. Fulton, *Algebraic Curves*, Benjamin/Cummings Publishing Company, Reading, Massachusetts, 1981.

[H1] I. N. Herstein, *Topics in Algebra*, John Wiley & Sons, New York, 1975.
[H2] T. W. Hungerford, *Algebra*, Springer-Verlag, New York, 1974.
[R] M. Reid, *Algebraic Geometry*, Oxford Press, Oxford, 1970.
[W] B. L. van der Wärden, *Algebra*, Frederick Ungar Publishing Co., New York, 1970.

Department of Mathematics
West Georgia College
Carrollton, GA 30118
mfaucett@westga.edu

To learn Calculus without understanding what led to its development and how it was used by Newton and others, is like learning to play scales on the piano without being shown any compositions.

—*F. J. Swertz*

The incorporation of history in the teaching of mathematics is essential if the ideas of its purpose, its structure, its wonder, its creativeness are to be aroused in the child.

—*F. J. Swertz*

History is commonly taught in schools to initiate the young into a community—to give them an awareness of tradition, a feeling of belonging, and a sense of participation in an ongoing process or institution. Similar goals can be advocated for the teaching of the history of mathematics.

—*F. J. Swertz*

NOTES

Edited by: **John Duncan**

Transitive Cellular Automata Are Sensitive

Bruno Codenotti and Luciano Margara

The notion of chaos is a very appealing one, and it has intrigued several scientists (see [**1**, **2**, **5**, **7**] for some work on the properties that characterize a chaotic process). There are simple deterministic dynamical systems that exhibit unpredictable behavior. Though counter-intuitive, this fact has a very clear explanation. The lack of infinite precision causes a loss of information which is dramatic for some processes which quickly lose their deterministic nature to assume a non deterministic (unpredictable) one. This observation leads to the intuition that investigations about chaos are intrinsically of an interdisciplinary nature. Indeed, the study of chaos draws its deeper methods of analysis from mathematics, it owes to physics a treasure of important problems, and it brings challenges to the science of computing. The reason for this last fact relies on the above observation that a chaotic behavior lies in between two different modes of computation, determinism and nondeterminism, whose quantitative comparison is central to the main open questions in the theory of computing [**6**].

A chaotic phenomenon can indeed be viewed as a deterministic one, in the presence of infinite precision, and as a nondeterministic one, in the presence of finite precision constraints (see Figure 1). Thus one should look at chaotic processes as at processes merged into time, space, and precision bounds, which are the key resources in the science of computing.

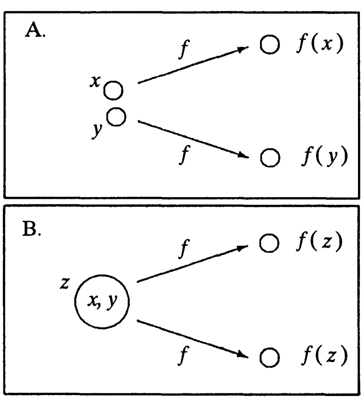


Figure 1. (A) f is deterministic. In the case of finite precision (B) x and y are not distinguishable and the map f loses its deterministic nature.

A nice way in which one can analyze this finite/infinite dichotomy is by using cellular automata models (CA). We skip for now the technical definitions, and proceed by giving a simple example of CA (see Figure 2).

```
                           Shift direction
                           <---------------

  x = ...?????????|010101110|??????????...
         σ^1(x) = |10101110?|
                  |0101110??|
                  |101110???|
                  |01110????|
             .    |1110?????|
             .    |110??????|
             .    |10???????|
                  |0????????|
         σ^9(x) = |?????????|
```

Figure 2. Finite precision combined with sensitivity to initial conditions causes unpredictability after a few iterations (x represents the state of the CA at time step 0, and $\sigma^i(x)$ the state at time step i).

Let $X = Z_2^Z$ and σ be the shift map on X. (X, σ) is a one-dimensional CA, where each element of Z_2 is contained in a *cell* of the CA and each cell can communicate with two neighbor cells. In order to completely describe the elements of X, we need to operate on sequences of binary digits of infinite length. Assume for a moment that this be possible. Then the shift map would be completely predictable, i.e., one can completely describe $\sigma^n(x)$, for any $x \in X$ and for any integer n.

In practice only finite objects can be computationally manipulated. Let $x \in X$. We assume to know a portion of x of length n (the portion between the two vertical lines in Figure 2). One can easily verify that $\sigma^n(x)$ completely depends on the unknown portion of x. In other words, if we have finite precision, the shift map becomes unpredictable, as a consequence of the combination of the finite precision representation of x and the sensitivity of σ.

CA can be used to investigate the nature of chaotic processes. What is nice about them is that they are dynamical systems as well as models of computation. In addition they are so simple as to allow a mathematical treatment.

We prove that for CA the sensitivity to initial conditions, which is widely recognized to play a key role in chaotic phenomena, is implied by transitivity. This result is important because it reveals the extent to which transitivity determines the behaviour of a CA. We also show that this is not always the case, by working out the example of subshifts of finite type. More precisely, the main technical contribution of this note is in the proof of the two following claims.

Claim 1. Transitive CA are sensitive to initial conditions.
Claim 2. A subshift of finite type $\mathscr{S}$ is transitive but not sensitive if and only if $\mathscr{S} = \{x, \sigma(x), \dots, \sigma^p(x)\}$ for some $x \in Z_2^Z$ and $p > 0$.

Basic definitions. Informally, CA are dynamical systems which consist of a regular lattice of variables (sometimes called cells or sites) which can assume a finite number of discrete values. The state of the CA, specified by the values of the variables at a given time, evolves in synchronous discrete time steps according to a given local rule. Hedlund ([4]) characterized the class of CA as follows. A map F defined on a m-dimensional regular lattice L is a CA if and only if it is continuous and it commutes with the m relevant shift maps defined on L.

Let X be a set and $f, f\colon X \to X$, be a map. We say that f is transitive if for all non empty open subsets U and V of X there exists a natural number n such that $f^n(U) \cap V \neq \varnothing$.

Let d be a distance on X. f is sensitive to initial conditions if there exists $\delta > 0$ such that for any $x \in X$ and for any neighborhood $N(x)$ of x, there is a point $y \in N(x)$ and a natural number n, such that $d(f^n(x), f^n(y)) > \delta$ (δ is called sensitivity constant). Note that if f is sensitive to initial conditions, then for all practical purposes, its dynamics defies numerical computation. In fact, due to round-off errors, the finite-precision computation of an orbit, no matter how accurate, may be very different from the real orbit.

We define the following notion of distance on Z_2^Z.

$$d(a,b) = \sum_{i=-\infty}^{+\infty} \frac{1}{2^{|i|}} |a(i) - b(i)|, \quad \forall a, b \in Z_2^Z. \tag{1}$$

It is easy to verify that d is a metric and that the metric topology induced by d coincides with the product topology induced by the discrete topology of Z_2. With this topology, Z_2^Z is a compact and totally disconnected space and the shift map σ is continuous.

Proof of Claim 1. (For simplicity, we give the proof for the case of CA defined on Z_2^Z. The complete proof of Claim 1 for CA defined on Z_k^L, where $L = Z^h$ can be found in [3].) Let r be an integer and $f, f\colon Z_2^{2r+1} \to Z_2$, be a map. An elementary CA is a pair (X, F), where $X = Z_2^Z$ and F is defined as follows.

$$F(c)(i) = f(c(i-r), \dots, c(i+r)), \quad \forall i \in Z.$$

Let (X, F) be a nonsensitive elementary CA. Then for each $\delta > 0$, there exist a configuration $c \in X$ and a neighborhood $N(c)$ of c, such that for any $c' \in N(c)$ and for any integer $n \geq 0$, $d(F^n(c), F^n(c')) \leq \delta$.

Let $c_{[h,k]}, h < k$, denote the sequence $\langle c(h), c(h+1), \dots, c(k) \rangle$. We define two constants p and q according to the value of δ and to $N(c)$, respectively. Let p be the smallest integer such that

$$(a_{[-p,p]} = b_{[-p,p]}) \Rightarrow (d(a,b) \leq \delta) \quad \forall a, b \in X.$$

Let q be the smallest integer such that

$$(a_{[-q,q]} = c_{[-q,q]}) \Rightarrow (a \in N(c)) \quad \forall a \in X.$$

One can easily verify that, for any $c' \in N(c)$ and for any integer n, $F^n(c')(0) = F^n(c)(0)$. In other words, $F^n(c')(0)$ does not depend on $c'(j), |j| > q$.

We define an open subset U of X as follows.

$$U = \{z \in X | z_{[-2q,0]} = c_{[-q,q]} \text{ and } z_{[1,2q+1]} = c_{[-q,q]}\}.$$

If $z \in U$, for any integer $n \geq 0$, we have

$$F^n(z)(-q) = F^n(c)(0) = F^n(z)(q+1).$$

Consider now the following open set

$$V = \{z \in X | z(-q) = 0 \text{ and } z(q+1) = 1\}.$$

Then we have that, $\forall n \geq 0, F^n(U) \cap V = \emptyset$. ∎

Proof of Claim 2. A subshift space $\mathcal{S}$ is any closed and shift invariant subset of Z_2^Z. Let $B = \{b_1, \ldots, b_n\}$ be any finite set of finite length binary blocks. We say that $\mathcal{S}$ is a subshift of finite type with forbidden blocks B if and only if $\mathcal{S}$ is the set of all the elements of Z_2^Z which contain none of the blocks from B.

Let $\mathcal{S}$ be a subshift of finite type with forbidden blocks $B = \{b_1, \ldots, b_n\}$. Without loss of generality, we can assume that each block has length m. $\mathcal{S}$ can be described by a finite directed graph $G_{\mathcal{S}} = \{V, E\}$. Each vertex of V is labelled by a block of length m which does not belong to B. Two vertices $v = \langle v_1, \ldots, v_m \rangle$ and $w = \langle w_1, \ldots, w_m \rangle$ are connected by an edge of E if and only if $v_2 = w_1, \ldots, v_m = w_{m-1}$. Every edge $\langle v_1, \ldots, v_m \rangle \to \langle w_1, \ldots, w_m \rangle$ of E is labelled by w_m. One can easily see that each element of $\mathcal{S}$ is given by a suitable sequence of labels read on a bi-infinite path on the graph $G_{\mathcal{S}}$. The converse is also true. Note that for any subshift of finite type $\mathcal{S}$ the two following properties hold.

- $\mathcal{S}$ is transitive if and only if for every ordered pair of vertices v and w there is a path in $G_{\mathcal{S}}$ starting at v and ending at w.
- $\mathcal{S}$ is transitive but not sensitive if and only if each vertex of $G_{\mathcal{S}}$ has exactly 1 outcoming edge.

These facts imply that $\mathcal{S}$ is a transitive but not sensitive subshift of finite type if and only if $\mathcal{S} = \{x, \sigma(x), \ldots, \sigma^p(x)\}$ for some $x \in Z_2^Z$ and $p > 0$. ∎

We believe that the analysis of chaotic processes is at the very heart of central problems in different sciences. We hope that this note, which has provided further results on the relation between chaos, transitivity, and sensitivity, could be interpreted as a modest contribution in this direction.

ACKNOWLEDGMENT. We wish to thank Paola Favati and Grazia Lotti for many useful discussions.

REFERENCES

1. D. Assaf, IV and W. A. Coppel, Definition of Chaos. *Amer. Math. Monthly*, 865, 1992.
2. J. Banks, J. Brooks, G. Cairns, G. Davis, and P. Stacey, On the Devaney's Definition of Chaos. *Amer. Math. Monthly*, 332–334, 1992.
3. L. Margara, For additive Cellular Automata, Transitivity Implies Sensitivity. *IMC Technical Report*, 1994.
4. G. A. Hedlund, Endomorphism and Automorphism of the Shift Dynamical System. *Mathematical System Theory* 3(4), 320–375, 1970.
5. C. Knudsen, Chaos Without Nonperiodicity. *Amer. Math. Monthly*, 563–565, 1994.

6. M. Sipser, The History and the Status of the P Versus NP Question. *24th Annual ACM Symposium on the Theory of Computing*, 603–618, 1992.
7. M. Vellekoop and R. Berglund, On Intervals, Transitivity = Chaos. *Amer. Math. Monthly*, 353–355, 1994.

Intituto di Matematica Computazionale del CNR
Pisa, ITALY
codenotti@iei.ip.cnr.it

Dipartimento di Informatica
University of Bologna
ITALY
and
Dipartimento di Informatica
University of Pisa
ITALY
margara@di.unipi.it

Use of a Differential Inequality in Combinatorics

Ray Redheffer

Throughout this note x, y, b are real, n and k are integers, $n \geq 2$, and

$$\binom{x}{n} = \frac{x(x-1)\cdots(x-n+1)}{n!}.$$

Let F be any continuous real-valued function satisfying

$$F(x) = 0 \Rightarrow F'(x) < 0 \tag{1}$$

on an open interval I containing b. If $F(b) = 0$, then $F(x) > 0$ for $x < b$ and $F(x) < 0$ for $x > b$. We will use this rather trivial observation to prove

Theorem 1. *With $y > n - 1$, $b > n - 1$ and $x > n - 2$, suppose*

$$\binom{y}{n} = \binom{b}{n} + \binom{x}{n-1}.$$

Then the inequality

$$\binom{y}{n+1} > \binom{b}{n+1} + \binom{x}{n}$$

holds if $b > x$ and the reversed inequality holds if $b < x$.

Theorem 1 gives a simple approach to some well-known theorems in combinatorics [**1**, **2**, **3**, **4**], and it is left for devotees of this field to explore that aspect of the subject on their own. Attention here is confined to the proof of Theorem 1, which involves a novel use of differential inequalities.

Proof: Since $y, b > n - 1$ and $x > n - 2$, the functions

$$Y = \binom{y}{n}, \qquad B = \binom{b}{n}, \qquad X = \binom{x}{n-1}$$

are positive and $Y = B + X$ by hypothesis. With b fixed, this yields $y = \phi(x)$ as a strictly increasing differentiable function. If

$$F(x) = \binom{y}{n+1} - \binom{b}{n+1} - \binom{x}{n},$$

it is to be shown that $F(x) > 0$ for $x < b$ and $F(x) < 0$ for $x > b$. By the Pascal equality $x = b \Rightarrow y = b + 1$, and by the Pascal equality again, these two conditions imply $F(b) = 0$. Hence the conclusion follows as soon as we establish (1).

It is easily checked that

$$F(x) = \frac{y-n}{n+1}Y - \frac{b-n}{n+1}B - \frac{x-n+1}{n}X. \tag{2}$$

The equation $Y = B + X$ shows that $y > b$ and also that

$$F(x) = \frac{y-n}{n+1}X + \frac{y-b}{n+1}B - \frac{x-n+1}{n}X.$$

Hence $F(x) = 0$ implies

$$\frac{y-n}{n+1} < \frac{x-n+1}{n}. \tag{3}$$

This completes the first step of the proof.

For the second step, the equation $Y = B + X$ gives $Y'y' = X'$ where $Y' = dY/dy$ and, aside from this, primes denote differentiation with respect to x. Hence by (2)

$$F'(x) = \frac{y-n}{n+1}X' + \frac{y'}{n+1}Y - \frac{x-n+1}{n}X' - \frac{1}{n}X.$$

Therefore $F'(x) < 0$ holds subject to (3) if $Yy'/(n+1) \le X/n$, or

$$\frac{1}{n+1}\frac{X'}{X} \le \frac{1}{n}\frac{Y'}{Y}. \tag{4}$$

This completes the second step.

Differentiating

$$\ln X = \sum_{k=1}^{n-1} \ln(x-k+1) - \ln(n-1)!, \quad \ln Y = \sum_{k=0}^{n-1} \ln(y-k) - \ln n!,$$

we see that (4) holds if

$$\frac{1}{n+1}\sum_{k=1}^{n-1}\frac{1}{x-k+1} \le \frac{1}{ny} + \frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{y-k}. \tag{5}$$

We now add the inequality $(n-k)/(n+1) < (n-k)/n$ to (3) to get

$$\frac{y-k}{n+1} < \frac{x-k+1}{n},$$

which is to say

$$\frac{1}{n+1}\frac{1}{x-k+1} < \frac{1}{n}\frac{1}{y-k}, \qquad 1 \le k \le n-1.$$

This gives (5) and completes the proof.

ACKNOWLEDGMENT. Theorem 1 is not due to me. It was brought to me by Prof. Basil Gordon with a request for a simple proof, and he also provided all references to the literature. The theorem and its proof were presented at the University of Karlsruhe in December 1991 under auspices of the Mathematical Institute I.

REFERENCES

1. P. Erdös, C. Ko and R. Rado, *Quart. J. Math. Oxford* II 12 1961, 313–320.
2. G. Katona, *Theory of graphs*, Akadémiai Kiadó 1966, 187–207.
3. J. B. Kruskal, *Math. Opt. Techniques*, Univ. Cal. Press 1963, 251–278.
4. L. Lovász, Combinatorial problems and exercises, North Holland Pub Co. 1979, 18, 81, 169–172, 458–461.

Department of Mathematics
UCLA
Los Angeles, CA 90024

Its use [i.e. the use of the history of mathematics] is not just that history may give everyone its due that others may look forward to similar praise, but also that the art of discovery be promoted and its method known through illustrious examples.

—*Leibniz*

Edited by **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Evolution of Methods of Convex Optimization

V. M. Tikhomirov

It is surprising that the story of convex optimization began only some fifty years ago. What is even more surprising is that the great analysts and specialists in the calculus of variations did not consider constraints in the form of inequalities. Before the 1940s there were very few papers devoted to the theory of inequalities. Of these, the most significant are a paper by Fourier (1823) and a paper by Vallée-Poussin (1911) (see [1]). Since that time, tens of thousands of papers have been devoted to this subject! A great many of them deal with convex problems.

The earliest nontrivial problems involving constraints in the form of inequalities appeared in a paper by L. V. Kantorovich. This happened as follows. In the spring of 1939 Kantorovich, then a young 26-year-old professor at Leningrad University, was approached by engineers of a plywood factory. They wanted to make more efficient use of their machine tools but lacked the background to deal with the mathematical version of their problem. The problem they put before the future recipient of a Nobel prize was very simple; in fact it was virtually a high school problem. But the young scholar treated it very seriously. Some 35 years later he wrote: "As it turned out, this was not a casual problem. I noticed a great many problems of different contents and of much the same mathematical character." They all fitted the scheme

$$\begin{array}{ccc} \langle c, x\rangle \to \sup, & & \sum_{i=1}^{n} c_i x_i \to \sup, \\ & \Leftrightarrow & \\ Ax \le b & & \sum_{i=1}^{n} a_{ji} x_i \le b_j,\ 1 \le j \le m \end{array} \tag{1}$$

and came to be known as linear programming problems.

In the same year, in 1939, Kantorovich wrote a monograph [2] devoted to methods of solution of problem (1). He tried to interest the Soviet authorities in his investigations because he thought that they could be of use for the development of the Soviet economy. But according to some ideological doctrine, an abstract subject like mathematics was of no conceivable use to so life-related a subject as economics. As a result, Kantorovich was rudely told to keep out of this range of issues.

The vigorous development of linear programming, and then of convex optimization in a wider sense, began in 1947 in the United States. The story of linear

programming was told by G. B. Dantzig in one of the papers in [3] devoted to the origins and evolution of mathematical programming. According to Dantzig, the swift progress of linear programming was due primarily to such eminent scholars as von Neumann, Kantorovich, Leont'ev and Koopmans. To these names one should definitely add Dantzig's own name. One of his greatest contributions was the development of a remarkable algorithm for the solution of linear programming problems known as

THE SIMPLEX METHOD. The essence of Dantzig's method is very simple. A careful look at problem (1) shows that the feasible vectors in this problem, that is vectors x for which $Ax \le b$, form a polyhedron in $\mathbf{R}^n$ (which may or may not be bounded; if bounded it is called a convex polytope). The polyhedron is the intersection of a finite number of halfspaces. A linear function on such a set (satisfying the additional assumption that it contains no straight lines) attains its maximum (if at all) at one of its vertices. This means that we need only look over the values of the linear function in (1) which is to be maximized on the set of vertices and choose the largest. However, in applied problems the number of vertices can reach astronomical proportions, hence the need for a systematic search. Such a search procedure was devised by Dantzig. What follows is a description of Dantzig's simplex method in the so-called *nondegenerate* case.

Suppose we know a vertex. (There are effective methods for finding vertices.) We shall assume that this vertex is *nondegenerate*. This means that at this vertex exactly n "linearly independent" inequalities become equalities. An explanation is in order.

Without loss of generality, we can assume that the first n inequalities become equalities, that is

$$\begin{aligned} &\langle a^j, \bar{x}\rangle = b_j,\ 1 \le j \le n,\ \langle a^j, \bar{x}\rangle < b_j,\ j \ge n+1, \\ &a^j := (a_{j1}, \dots, a_{jn}),\ j = 1, \dots, m. \end{aligned} \tag{i}$$

Nondegeneracy of the vertex means that the vectors $\{a^1, \dots, a^n\}$ form a basis of $\mathbf{R}^n$, that is, the matrix $A_n := (a_{ij})_{1 \le i, j \le n}$ is nonsingular. Put $\bar{b} = (b_1, \dots, b_n) \in \mathbf{R}^n$. Then

$$A_n\bar{x} = \bar{b} = (b_1, \dots, b_n) \Leftrightarrow \bar{x} = A_n^{-1}\bar{b}. \tag{ii}$$

We solve the equation

$$A_n^T\bar{\lambda} = c \tag{iii}$$

(A_n^T denotes the transpose of A_n.) We have the alternative: (I) $\bar{\lambda} = (\lambda_1, \dots, \lambda_n) \ge 0$ or (II) some of the components of $\bar{\lambda}$ are negative.

We shall show that in the first case $\bar{x}$ is a solution of (1). In fact, let x be a feasible vector ($Ax \le b$ (iv)). Put $\lambda = (\lambda_1, \dots, \lambda_n, 0 \dots 0)$. Then

$$\begin{aligned} &\langle c, x\rangle \overset{(iii)}{=} \langle A_n^T\bar{\lambda}, x\rangle \overset{Id}{=} \langle \bar{\lambda}, A_n x\rangle \overset{Id}{=} \langle \lambda, Ax\rangle \overset{(iv) \text{ and } \lambda \ge 0}{\le} \\ &\le \langle \lambda, b\rangle \overset{Id}{=} \langle \bar{\lambda}, \bar{b}\rangle \overset{(ii)}{=} \langle \bar{\lambda}, A_n\bar{x}\rangle \overset{Id}{=} \langle A_n^T\bar{\lambda}, \bar{x}\rangle \overset{(iii)}{=} \langle c, \bar{x}\rangle, \end{aligned}$$

which was to be shown. We consider next the alternative possibility.

Suppose that, say, $\lambda_1 < 0$. We find a nontrivial solution of the homogeneous system

$$\langle a^2, y\rangle = \dots = \langle a^n, y\rangle = 0. \tag{v}$$

In view of the nonsingular nature of A_n we find that $\langle a^1, y\rangle \neq 0$. With possibly a change of sign we have

$$\langle a^1, y\rangle = -\varepsilon < 0. \qquad (vi)$$

Then for small values of $t > 0$ we have

$$\langle a^j, \bar{x} + ty\rangle < b_j,\ j = 1,\ j \geq n+1,\ \langle a^j, \bar{x} + ty\rangle = b_j,\ 2 \leq j \leq n,$$

that is the vector $\bar{x} + ty$ is feasible. Also,

$$\langle c, \bar{x} + ty\rangle \overset{(v),(vi)}{=} \langle c, \bar{x}\rangle + t\langle c, A_n^{-1}(-\varepsilon, 0, \ldots, 0)\rangle \overset{Id}{=} \langle c, \bar{x}\rangle$$
$$+ t\left\langle \left(A_n^{-1}\right)^T c, (-\varepsilon, 0, \ldots, 0)\right\rangle \overset{Id}{=} \langle c, \bar{x}\rangle - t\varepsilon\lambda_1.$$

This means that $\langle c, \bar{x} + ty\rangle > \langle c, \bar{x}\rangle$ for all $t > 0$.

If $\bar{x} + ty$ is feasible for all $t > 0$ then the supremum for the problem is $+\infty$. Otherwise, $\langle a^j, \bar{x} + t_0 y\rangle = b_j$ for some value t_0 of t and for some $j \geq n+1$. Then $\bar{x} + t_0 y$ will take the place of $\bar{x}$ and we can perform another iteration. And so on. This completes the description of one version of the nonsingular nondegenerate simplex method.

The simplex method has played an important role in the history of numerical methods of optimization. For a long time it was not known whether or not the problem in (1) is *polynomial*, that is, whether or not there are for this problem algorithms that require, in all cases, just a polynomial number of operations (in terms of the size of the input). In 1970 Klee and Minty constructed examples which showed that in some situations the simplex method requires an exponential number of steps. But for some reason problems of this kind don't turn up in applications! Many mathematicians (including Dantzig) have said that it is something of a miracle that the simplex method has worked essentially without a hitch for about 50 years in countless applied situations. Dantzig himself said: "The tremendous power of the simplex method is a constant surprise to me."

Some years after his invention of the simplex method Dantzig decided to write a comprehensive monograph (in addition to the book [1]) devoted to linear programming. It was to be a survey of all papers devoted to inequalities and their applications to extremal problems. In particular, the monograph [2] of Kantorovich attracted his attention. Dantzig thought highly of the monograph and saw to it that it was translated. In this way the scientific world learned about Kantorovich's—and Koopman's—contributions to the theory. In 1975 the two scientists were awarded the Nobel prize in economics for their contribution to the development of linear programming and for its application to economics. Next we discuss other approaches to linear and convex optimization. We begin with a description of

THE METHOD OF CENTRAL SECTIONS. I was a witness to the story I am about to tell. In 1962 I worked in Voronezh, a big town in central Russia. At that time, the Voronezh school of mathematics, headed by M. A. Krasnosel'ski, flourished. The school pursued important scientific objectives and tried to enter various applied areas. In particular, in the early sixties, Krasnosel'ski signed a contract with an applied mathematics firm headed by David B. Yudin. Yudin's name was known to specialists in optimization from the monograph [4] written by him and his then coworker E. G. Gol'shtein. Yudin posed the following problem: *Find an effective algorithm that minimizes a sum of exponentials (with positive weights) on a compact polyhedron.*

The problem attracted the attention of Anatoli Yu. Levin, one of the main heroes of our story, and a member of Krasnosel'ski's group. Levin pondered this, as well as more general problems, for a long time. He discussed them with colleagues, including myself, and one day hit on a remarkable idea for an algorithm that turned out to be applicable to the following more general problem: *find the minimum of a convex (as well as a quasiconvex) function f on a finite-dimensional convex body A*, i.e.

$$f(x) \to \inf;\ x \in A.$$

This is, essentially, the general problem of convex optimization. What follows is a description of Levin's algorithm for a smooth convex function.

We denote A by A_1. We determine $x_1 = grA_1$, the center of gravity of A_1. Then we compute $f'(x_1)$. If this is the zero vector, then the problem is solved. Otherwise we eliminate the part of A_1 in the halfspace $\Pi_1' := \{x | \langle f'(x_1), x - x_1 \rangle > 0\}$. (This step is justified as follows: For a convex f it is easy to show that $f(x) - f(x_1) \geq \langle f'(x_1), x - x_1 \rangle$. But then for $x \in A_1 \cap \Pi_1'$ we have $f(x) > f(x_1) > \min$.) We denote the remaining part of A_1 by A_2 and repeat the procedure. And so on.

If we take as ξ_m a point of $\{x_1, \dots, x_m\}$ at which $f(\xi_m)$ is not less than any of the values $f(x_i)$, $1 \leq i \leq m$, then it can be shown that $f(\xi_m)$ tends to the minimum of f on A and the error in the value of f decreases at the rate of a geometric progression. Also, the volume of A_m decreases exponentially. This fact is a consequence of the following result in convex geometry due to Grünbaum: *Let A be a convex body in $\boldsymbol{R}^n$, and let $\xi = grA$ be its center of gravity. Every hyperplane passing through ξ divides A into two parts such that the volume of each of these parts is no less than the fraction* $1 - (1/e)$ *of the volume of A.*

We note that in the one-dimensional case the method just described reduces to halving of intervals.

Levin delayed the writing up of his result for publication. (Some of the reasons for this delay will be given below.) In the meantime the American D. J. Newman independently hit on the idea of the method of central sections. The Levin and Newman papers [**5**, **6**] appeared simultaneously in 1965.

Some time later, my student A. I. Kuzovkin and I supplemented Levin's algorithm and showed that it is possible to obtain an exponential rate when computing the value of f rather its gradient [**7**].

After that the issue was dormant. The next important event occurred some 15 years after the discovery of the method of central sections.

THE METHOD OF CIRCUMSCRIBED ELLIPSOIDS OF NEMIROVSKI-YUDIN-SHOR. Let me introduce to the reader one more person who played an important role in the subsequent evolution of convex optimization. This is Arkadi Nemirovski, a graduate of Moscow University and one of the last students of G. E. Shilov.

After the completion of graduate studies in 1974 Nemirovski began to work for Yudin. Yudin put before his young coworker the issue of the complexity of the solution of the problem of convex optimization. In November of 1974, (while walking in the forest) Nemirovski hit on yet another method of solution of the problem of convex optimization. It came to be known as the method of circumscribed ellipsoids. The method was presented in a paper. The key idea of this method was found independently (and somewhat later) by the well-known special-

ist on convex optimization, the Kiev mathematician Naum Z. Shor; see [18]. That is why the method of circumscribed ellipsoids is sometimes referred to as the *Nemirovski-Yudin-Shor method.*

The method combines two ideas. One is the method of sections discussed above and the other is the geometric fact that *half an ellipsoid can be put in an ellipsoid of smaller volume than the initial ellipsoid.*

We now describe in greater detail the Nemirovski-Yudin-Shor algorithm. We denote by E_0 an ellipsoid circumscribed about A. If the center c_0 of this ellipsoid lies outside A, then we pass through it a halfspace not containing points of A and eliminate the half of the ellipsoid that doesn't intersect A. If $c_0 \in A$, then we compute $f'(c_0)$, carry out a section à la Levin-Newman,, and again end up with half an ellipsoid which we denote by E_0'. Next we circumscribe about E_0' an ellipsoid of smaller volume than the volume of E_0', denote it by E_1, and repeat our procedure.

The rate of the solution just described is that of a geometric progression. Yudin and Nemirovski showed in their paper that the Levin-Newman method of central sections cannot be substantially improved in the class of convergent algorithms of minimization of convex functions. The method of circumscribed ellipsoids is somewhat inferior to the method of central sections in terms of the rate of convergence but has the advantage that it obviates the need for finding centers of gravity of polyhedrons.

All this was told by Nemirovski and Yudin at various seminars. Once they described this method at the seminar of E. G. Gol'shtein in the Central Economics-Mathematics Institute. Gol'shtein liked the Nemirovski-Yudin lecture very much but observed that similar ideas had been advanced earlier. "Where?" asked the lecturers. "In A. Yu. Levin's paper" was the answer. "Which Levin?" asked Yudin nervously. "The same Levin who solved your problem 15 years ago..." replied Gol'shtein. Accordingly, Nemirovski and Yudin included in their paper [8] a reference to A. Yu. Levin's paper [5].

THE PAPER OF L. G. KHACHIAN. A few years passed during which only one event occurred that bears on our story. L. Levin, Nemirovski's one time fellow-student, decided to emigrate to America and Nemirovski went to bid him farewell. He brought with him a copy of his paper [8] on the method of ellipsoids and gave it to Levin. At the time Levin was not in a mood for mathematics and put the paper at the bottom of his valise.

One early morning I happened to be passing the Computing Center of the Academy of Sciences in Moscow and noticed the racing figure of a former fellow-student of mine who worked at the time at the Computer Center. I exclaimed: "What happened?" Without slowing down he shouted incoherently "Khachian... New York Times... Press conference... I am late... ."

This is what happened. A young associate at the Computing Center, Leonid G. Khachian, had published a note [9] titled "A polynomial algorithm in linear programming" in the Proceedings of the Soviet Academy of Sciences. No one reacted to this paper for a long time. But then a conference of specialists in the area of convex optimization took place in the U.S. One of the participants presented a survey of recent progress in the area and mentioned the paper of Khachian. While describing the latter, he remarked: "Khachian shows that the linear programming problem can be solved in polynomial time... ." He was about to continue his lecture but was interrupted by the Hungarian mathematician Peter

Gács: "Would you mind repeating what you just said?" The lecturer obliged. Gács exclaimed: "But this is the solution of a famous problem, the one about the polynomial nature of the problem of linear programming!" It turned out that Khachian's result could indeed be interpreted as a justification of the claim of the polynomial nature of the problem of linear programming. The algorithm turned out to be an application of the method of ellipsoids to the problem of linear programming. The press got interested in Khachian's paper and this led to a The New York Times press conference in the Computing Center of the Soviet Academy of Sciences... .

Khachian's paper contained a reference to Shor's paper. This reminded L. Levin of the copy that Nemirovski had given him at the time of his departure to the U.S. From it he learned about the method of circumscribed ellipsoids which yielded a polynomial estimate for the number of steps in the solution of the problem of convex programming; in particular, for linear programming over the reals. It was Khachian who took the next step, essential for everything that was to follow, namely, he found the necessary computational complexity for the rationals (and it turned out to be polynomial).

As a result of all this the papers of Khachian and Nemirovski-Yudin became widely known. Their authors were honored with the prestigious Fulkerson award of the International Society for Mathematical Programming and the American Mathematical Society for 1982.

Peter Gács called L. Levin's attention to the paper [5] of his namesake A. Yu. Levin. Now we must go back somewhat in time. One of the reasons for the delay in the publication of the paper [5] was that A. Yu. Levin wanted to overcome the difficulties associated with the finding of the center of gravity of a polyhedron. (It later turned out that problem has indeed exponential complexity.) In [5] Levin advanced the idea that it was sometimes necessary to circumscribe simplexes about the A_n, and that this could be done without impairing the exponential decrease of their volumes. He gave no specific algorithms. Rather, it was an existence theorem for such algorithms. Beginning with this idea, Levin and his coauthor Boris Yamnitsky showed in [10] that A. Yu. Levin's idea was correct in principle. The paper [10] contains a description of the method of "circumscribed simplexes" which coincides with the method of circumscribed ellipsoids, runs in polynomial time, and is in some respects superior to the method of ellipsoids.

CONCLUDING REMARKS. The events just described have resulted in an explosive development of methods of convex optimization in the eighties. There appeared a number of methods of sections (such as the method of inscribed ellipsoids of Tarasov-Khachian-Ehrlich, (see [19]) and others). Many methods were modified and perfected. Also, in 1984 Karmarkar proposed a polynomial method for the solution of problems of linear programming based on different ideas. This method turned out to offer many advantages and it resulted in a veritable flood of papers on algorithms dealing with convex programming in which the methods discussed above were modified, developed and perfected. The reader can find the relevant details in the monograph [11] of Nemirovski and Nesterov.

For additional information in English, see items [12]–[19].

ACKNOWLEDGMENT. The column editor wishes to thank Professor Michael J. Todd for his help.

REFERENCES

1. G. B. Dantzig, Linear Programming and Extensions, Princeton, Univ. Press, 1963.
2. Kantorovich, L. V. Mathematical methods in the organization and planning of production. Leningrad University, 1939. [Russian] For an English version see Management Science, Vol. 6, 1960, 366–422.
3. History of Mathematical Programming (A Collection of Personal Reminiscences). Ed. by J. K. Lenstra, F. H. G. Rinnooy Kan, A. Schrijver, North-Holland, 1991.
4. Yudin, D. B. and Gol'shtein, E. G. Problems and methods of linear programming. Moscow. Sovradio, 1961. [Russian]
5. Levin, A. Yu. On an algorithm for the minimization of convex functions. DAN USSR vol. 160, issue 6, 1965, pp. 1244–1247. [Russian]
6. Newman, D. J. Location of maximum on unimodal surfaces. Journ. of the Assoc. for Computing Machinery 12, 1965, pp. 395–398.
7. Kuzovkin, A. I. and Tikhomirov, V. M. On the number of computations needed for finding the minimum of a convex function. Economics and mathematical methods, 3:1, 1967, pp. 95–103. [Russian]
8. Yudin, D. B. and Nemirovski, A. S. Informational complexity and effective methods of solution of convex extremal problems. Economics and mathematical methods vol. 12, issue 1, 1976, pp. 357–369. [Russian]
9. Khachian, L. G. A polynomial algorithm in linear programming. DAN USSR vol. 244, issue 5, 1979, pp. 1093–1096 [Russian]. For an English version see Soviet Math. Doklady 20, 1979, 191–194.
10. Yamnitsky, B., Levin, L. An Old Linear Programming Algorithm Runs in Polynomial Time. In 23rd Annual Symposium on Foundations of Computer Science, IEEE, New York, 1982, pp. 327–328.
11. Nemirovsky, A., Nesterov, Yu. Interior-Point Polynomial Algorithms in Convex programming. SIAM Studies in Applied Math., 1994.
12. Nemirovsky, A. S., Yudin, D. B. Problem Complexity and Method Efficiency in Optimization. John Wiley and Sons, Chichester, 1983.
13. Grötschel, M., Lovász, L., Schrijver, A. Geometric Algorithms and Combinatorial Optimization. Springer-Verlag, 1988.
14. Schrijver, A. Theory of Linear and Integer Programming. John Wiley and Sons, 1986.
15. Shor, N. Z. Minimization Methods for Nondifferentiable Functions. Springer-Verlag, Berlin, 1985.
16. Bland, R. G., Goldfarb, D., Todd, M. J. The ellipsoid method: a survey. Operations Research, 29 (1981), 1039–1091.
17. Goldfarb, D., Todd, M. J. Linear programming in "Handbooks in Operations Research and Management Science," vol. 1, Optimization (G. L. Nemhauser, A. H. G. Rinnooy Kan and M. J. Todd, eds.), North Holland, Amsterdam, 1989, 73–170.
18. Shor, N. Z. Cut-off method with space extension in convex programming problems, Cybernetics, 13 (1977), 94–96.
19. Tarasov, S. P., Khachiyan, L. G., Erlikh, I. I. The method of inscribed ellipsoids. Soviet Math. Doklady, 37 (1988), 226–230.

Department of Mathematics
Moscow State University
117234 Moscow, RUSSIA

An expert is someone who knows some of the worst mistakes that can be made in his subject, and how to avoid them.

—*Werner Heisenberg (1901–1976)*

Physics and Beyond. New York: Harper and Row, 1971

THE AUTHORS

DAVID FOWLER was an undergraduate and graduate student at Cambridge. After working at Manchester, he went in 1967 to Warwick, where he is now a Reader. He made the English translation of René Thom's *Structural Stability and Morphogenesis* (1975), and is currently working on a sequel to his *The Mathematics of Plato's Academy* (1987), developing further what a colleague has called his "long infatuation with the completely unknown period of Greek mathematics." He also enjoys struggling with John Wallis' dense prose; see his 'An approximation technique... in Wallis and Taylor', *Archive for History of Exact Sciences* 41 (1991) 189–233.

JIM HENLE, A. B. Dartmouth College, Ph.D., M.I.T., teaches at Smith College, writes papers in set theory, and plays second clarinet. Convinced that he understands now the nature of the communion between mathematics and music, he wonders why he is attracted to atonal mathematics and at the same time repelled by atonal music. He wonders also about the strong, but little-noticed link between love of mathematics and early training on a soprano woodwind.

BRUCE SOLOMON first studied Differential Geometry under Barrett O'Neill while an undergraduate at UCLA in 1977. He got a Ph.D. in the subject at Princeton (1982), and now practices (and preaches) it in the Math Department of Indiana University, Bloomington.

THE BORWEINS All male **Borweins** in Canada (and there are very few in the rest of the world) are mathematicians and are closely related. David is an emeritus professor at the University of Western Ontario. Sons, Jonathan and Peter, are both professors of mathematics at Simon Fraser University. They have studied or worked at many of Canada's major universities and have published more than twice as many papers as the total of their ages (154). While papers have been co-authored by every other possible combination of mathematical Borweins this is their first family paper.

ROLAND GIRGENSOHN came to Canada after he received his Ph.D. from Technische Universitaet Clausthal (Germany) in 1992. He met Jonathan Borwein at the University of Waterloo and has since worked with several different subsets of the three Borweins. He has just finished a post-doctoral fellowship at Simon Fraser University and is now at the University of Western Ontario. His main (but not only) mathematieal interest are non-differentiable and other "pathological" functions of analysis.

MARK FAUCETTE received both his bachelor's degree and master's degree from The University of Georgia in 1983. He attended Brown University, completing his doctorate in 1988. It was while a graduate student at Brown that he received his introduction to algebraic geometry from a course given by Joe Harris. After teaching five years in the Midwest, Professor Faucette returned to his native Atlanta, Georgia, where he now serves on the faculty of West Georgia College. His research interest is algebraic geometry.

MICHAEL DEAKIN is a senior lecturer in Applied Mathematics at Monash University. His interests lie in Biomathematics and the History of Mathematics. He edits the History column of the journal *Function*, a school mathematics magazine. It was a reader's request for information that led to his interest in Hypatia (and his MONTHLY article of March 1994).

JOAN HART will complete her Ph.D. in Mathematics in the spring of 1996 under the direction of Ken Kunen at the University of Wisconsin–Madison. In between earning a B.S. and an M.A. in Mathematics at Miami University in Oxford, Ohio, she worked for four years as a telecommunications software engineer at ROLM Corporation in Santa Clara, California. Her research interests are logic, set theory, set theoretic topology, and automated reasoning. Her favorite players are Glenn Gould and Albert Belle.

KEN KUNEN grew up in New York many years ago, learning how to count cookies at his mother's knee. He went on to transfinite cookie jars, receiving his Ph.D. in set theory at Stanford in 1965, under the direction of Dana Scott. Since then, he has been teaching at the University of Wisconsin–Madison. His current research interests include set theoretic topology, measure theory, and automated reasoning. His hobbies include set theory, cross-country skiing, and surfing the internet.

> He who seeks for methods without having a definite problem in mind seeks for the most part in vain.
>
> —*David Hilbert (1862–1943)*
>
> *Mathematical Problems.* Bulletin of the American Mathematical Society, Vol. 8, pp. 444–445.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before June 30, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10494. *Proposed by WMC Problems Group, Western Maryland College, Westminster, MD.*

For each positive integer n, evaluate the sum

$$\sum_{k=0}^{2n}(-1)^k\binom{4n}{2k} \Big/ \binom{2n}{k}.$$

10495. *Proposed by Dennis Spellman, Philadelphia, PA, and David Joyner and William P. Wardlaw, United States Naval Academy, Annapolis, MD.*

Let R be a *principal ideal ring*; that is, R is a commutative ring with 1 in which every ideal is of the form Ra for some $a \in R$. Prove or give a counterexample: If $a, b \in R$ are multiples of one another, then they are *unit* multiples of one another (that is, there is an invertible element $u \in R$ such that $a = ub$).

10496. *Proposed by Robert A. Russell, New York, NY.*

Let C_m^n denote the number of cells in an n dimensional polyomino formed by adding m *coats*, as described below, to a *monomino* (one-celled polyomino). A coat consists of just enough cells to cover each previously exposed $n-1$ dimensional cell face. Thus $C_0^n = 1$, $C_1^n = 2n+1$, and $C_2^n = 2n^2+2n+1$. Show that $C_m^n = C_n^m$.

10497. *Proposed by Klaus Huber, Darmstadt, Germany.*

The *Gaussian integers* are those complex numbers $x+iy$ for which x and y are integers. Given a complex number z, let $[z]$ denote the closest Gaussian integer to z, let z^* denote the complex conjugate of z, and let $N(z) = zz^*$. It is known that, if p is a rational prime with $p \equiv 1 \pmod 4$, then $p = a^2+b^2$ with integer a and b in an essentially unique way, and hence $p = \pi\pi^*$ with π a Gaussian integer in an essentially unique way. Reduction modulo π is defined by

$$\gamma \bmod \pi = \gamma - \left[\frac{\gamma\cdot\pi^*}{\pi\cdot\pi^*}\right]\cdot\pi.$$

A *reduced set of residues* $\{\alpha_i : i = 1\dots p-1\}$ modulo the Gaussian integer π can be defined by choosing g to be a primitive root modulo p and setting $\alpha_i = g^i \bmod \pi$. Show that

$$\sum_{i=1}^{p-1} N(\alpha_i) = \frac{p^2-1}{6}.$$

10498. *Proposed by Ray Redheffer, University of California, Los Angeles, CA.*

Consider the system of differential equations

$$\frac{dx}{dt} = -\big(x + a(t)y\big) \qquad \frac{dy}{dt} = -\big(b(t)x + y\big) \tag{$*$}$$

where $a(t)$ and $b(t)$ are positive, continuous and bounded for $0 \le t < \infty$.

If $\big(\sup a(t)\big)\big(\sup b(t)\big) < 1$, it is easy to prove that all solutions of $(*)$ tend to 0 as $t \to \infty$. Does the same conclusion follow if one assumes only that $\sup\big(a(t)b(t)\big) < 1$?

10499. *Proposed by David Day, University of Kentucky, Lexington, KY, and Ren-Cang Li, University of California, Berkeley, CA.*

Let $M = T + \operatorname{diag}(\alpha_i)$, where T is Hermitian Toeplitz and $\alpha_1,\dots,\alpha_n$ are real numbers with $\alpha_1 < \cdots < \alpha_n$. Let $\lambda_1 \le \cdots \le \lambda_n$ denote the eigenvalues of M. Show that

$$\min_{1\le i\le n-1}(\lambda_{i+1}-\lambda_i) \ge \min_{1\le i\le n-1}(\alpha_{i+1}-\alpha_i).$$

10500. *Proposed by Jeffrey C. Lagarias and Peter W. Shor, AT&T Bell Laboratories, Murray Hill, NJ.*

Consider the following three properties that a sequence $\{f(n) : n = 1,2,\dots\}$ of real numbers may have.

(P1) The sequence $\{f(n) : n = 1,2,\dots\}$ is bounded.

(P2) For each real $\lambda > 1$, the subsequence $\{f(\lfloor 2^{\lambda^n}\rfloor) : n = 1,2,\dots\}$ is bounded.

(P3) For each real $\lambda > 1$, the subsequence $\{f(\lfloor \lambda^{2^n}\rfloor) : n = 1,2,\dots\}$ is bounded.

Obviously (P1) $\Longrightarrow$ (P2) and (P1) $\Longrightarrow$ (P3). What other implications hold, if any?

NOTES

(**10499**) Definitions may be found in books on Matrix Theory such as Horn & Johnson, *Matrix Analysis.*

SOLUTIONS

K-linear Maps that Permute Polynomial Roots

10266 [1992, 957]. *Proposed by Daniel Shapiro and Patrick Rabau, The Ohio State University, Columbus, OH.*

Let L/K be an algebraic extension of fields and let $T: L \to L$ be a K-linear map. Then T will be said to have "property G" if, for each polynomial $f \in K[x]$, the set of roots of f that lie in L are permuted by T.

(a) If K is infinite, show that T with property G must be a K-automorphism of L.

(b) Determine all examples of T with property G that are not K-automorphisms of L.

(c) What happens if L/K is an infinite algebraic extension?

Solution by the proposers. We prove that K-linear maps with property G that are not K-automorphisms exist only when $K = \mathbf{F}_2$ and $L \in \{\mathbf{F}_8, \mathbf{F}_{16}, \mathbf{F}_{32}\}$, where $\mathbf{F}_q$ denotes the finite field with q elements. Let $T: L \to L$ be a K-linear map with property G, but which is not a K-automorphism. It is easy to check that T is bijective and fixes every elements of K.

We first consider the case where $n = [L : K]$ is finite. Let $\bar{K}$ be an algebraic closure of K that contains L. Choose $\beta_1, \dots, \beta_r \in L$ such that $L = K[\beta_1, \dots, \beta_r]$. For each i, let $m_i(x)$ be the minimum polynomial of β_i over K. For $f(x) = m_1(x) \cdots m_r(x)$, let $\tilde{L}$ be a splitting field of $f(x)$ over K with $K \subset \tilde{L} \subset \bar{K}$. Finally, Γ will denote $\mathrm{Gal}(\tilde{L}/K)$, the Galois group of $\tilde{L}$ over K. We shall refer to Joseph Rotman, *Galois Theory*, Springer-Verlag, 1990 for standard results of Galois theory. Such citations will be denoted by [**R**] followed by the particulars of the reference.

We first claim that if $\alpha \in L$, there exists $\sigma \in \Gamma$ such that $T(\alpha) = \sigma(\alpha)$. To prove this, let $g(x) \in K[x]$ be the minimum polynomial of α over K. By assumption $T(\alpha) \in L$ is a root of $g(x)$. Therefore, both $K(\alpha)$ and $K\big(T(\alpha)\big)$ are isomorphic to $K[x]/(g(x))$. It follows that there is an isomorphism $\theta: K(\alpha) \to K\big(T(\alpha)\big)$ with $\theta(\alpha) = T(\alpha)$ ([**R**], Corollary 29, page 40). Since $T(\alpha) \in L$, we see that $\tilde{L}$ is both a splitting field for $f(x)$ over $K(\alpha)$ and a splitting field for $f(x)$ over $K\big(T(\alpha)\big)$. Thus θ can be extended to an isomorphism $\sigma: \tilde{L} \to \tilde{L}$ with $\sigma(\alpha) = T(\alpha)$ ([**R**], Theorem 32, page 32).

For $\sigma \in \Gamma$, let $P_\sigma = \{\alpha \in L : T(\alpha) = \sigma(\alpha)\}$. Since both T and σ can be viewed as linear transformations of L over K, P_σ is a subspace of L. The results of the previous paragraph show that L is the union of the subspaces P_σ, $\sigma \in \Gamma$. As Γ is a finite set, we have a finite union.

To such a union we will apply the following

which must be set-wise stabilized by S. However, $S(e_0+e_1+e_3)=e_0+e_{-1}+e_{-3}$ is not in this set. This contradiction implies T is a K-automorphism if $[L:K]$ is finite and either $|K|>2$ or $[L:K]\notin\{3,4,5\}$.

Suppose $[L:K]$ is infinite. Since T is assumed to be a K-linear map with property G which is not a K-automorphism, there exists $\alpha,\beta\in L$ with $T(\alpha\beta)\neq T(\alpha)T(\beta)$. Since α,β are algebraic over K, there exists a field E with $\alpha,\beta\in E$, $K\subset E\subset L$, and $[E:K]\geq 6$. Let $\bar{L}$ be an algebraic closure of K which contains L. Suppose $\tilde{E}\subset\bar{L}$ is a splitting field of a polynomial whose roots generate E over K. Let $\gamma\in\tilde{E}$. If $m(x)$ be the minimum polynomial of γ over K, $m(x)$ splits into linear factors over $\tilde{E}$. But T has property G and thus $T(\gamma)$ is a root of $m(x)$. Therefore, $T(\gamma)\in\tilde{E}$. Obviously, $T(\gamma)\in L$. Thus, T set-wise fixes $E_1=\tilde{E}\cap L$. It follows that T has property G with respect to E_1/K. But $[E_1:K]\geq 6$ so, by the preceding argument, $T_{|E_1}$ is a K-automorphism, a contradiction. Therefore, maps with property G are K-automorphisms with the possible exceptions listed.

To see that these exceptions occur, let $K=\mathbf{F}_2$ and $L\in\{\mathbf{F}_8,\mathbf{F}_{16},\mathbf{F}_{32}\}$ with e_i as above. Consider the K-linear map which satisfies $R(e_i)=e_{-i}$ for all i. A straightforward enumeration of the orbits of the action of $\mathrm{Gal}(L/K)$ shows R set-wise fixes each orbit. Let $f\in K[x]$ be a polynomial with a root in L. Since L is a splitting field for L/K, each irreducible factor of f over K either splits completely or is irreducible over L. Those factors which split have linear factors whose roots form an orbit of the action of the Galois group. Since R fixes these orbits, R has property G. Finally, the map R is not a K-automorphism since it fixes e_0 while the only K-automorphism which fixes e_0 is the identity.

No other solutions were received.

An Identity of Jacobi

10279 [1993, 76]. *Proposed by M. Al-Ahmar, Al-Fateh University, Tripoli, Libya.*

Let k and n be integers with $0<k<n$, and let A be a real n by n orthogonal matrix with determinant 1. Let B be the upper left k by k submatrix of A, and let C be the lower right $(n-k)$ by $(n-k)$ submatrix of A.

(**a**) Show that $\det(B)=\det(C)$.

(**b**) Give a geometrical interpretation.

(**c**) Generalize to the case in which A is a unitary matrix.

Editorial comment. Some readers recognized parts (**a**) and (**c**) as special cases of Jacobi's identity for determinants of submatrices. This result is mentioned in F. R. Gantmacher, *The Theory of Matrices*, Chelsea, 1960, eq. 33, p. 21, and in Roger A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge, 1985, sect. 0.8.4, p. 21. For convenience, we give the result here.

Jacobi's Identity. *Let*

$$A=\begin{pmatrix} B & D\\ E & C\end{pmatrix}\qquad A^{-1}=\begin{pmatrix} W & X\\ Y & Z\end{pmatrix}$$

where B and W are k by k matrices. Then

$$(\det Z)(\det A)=\det B.$$

Proof. Equate the determinants of the two sides of the block-matrix equation

$$\begin{pmatrix} B & D\\ E & C\end{pmatrix}\begin{pmatrix} I & X\\ O & Z\end{pmatrix}=\begin{pmatrix} B & O\\ E & I\end{pmatrix}\tag{$*$}$$

where I stands for the identity matrix and O for the zero matrix of the appropriate size.

We have $\det A = 1$, so $\det Z = \det B$. In **(a)**, $Z = C^{\mathrm{T}}$, so $\det B = \det C$, and in **(c)**, $Z = \bar{C}^{\mathrm{T}}$, so $\det B = \overline{\det C}$.

All readers giving more than a reference followed the above method of proof. A geometric interpretation, as requested in **(b)**, can be given by looking at the the columns of the matrices in $(*)$. Those of the second factor on the left consist of k vectors from the standard basis and $n-k$ from the columns of the orthogonal matrix A^{-1}. The parallelepiped with these vectors as edges is spanned by a k dimensional cube and an $n-k$ dimensional cube. The columns of the matrix on the right, being the transform by an orthogonal matrix, has the same geometry. The proof amounts to comparing the two ways of finding the volume of such a parallelepiped by finding the volume of the figure obtained by projecting one set of orthonormal edges on the orthogonal complement of the other set. The corresponding question for *simplexes* was proposed by Murray S. Klamkin in Problem 1465, *Crux Mathematicorum* [1989, 207; 1990, 284].

Solved by D. Callan & G. Wahba, R. J. Chapman (U. K.), K. S. Kedlaya (student), M. K. Kinyon, M. S. Klamkin (Canada), A. D. Melas (Greece), J. K Merikoski (Finland), A. Nijenhuis, M. Qian, F. Schmidt, J. C. Vera Lizcano & J. Madroñero Pabón (students, Colombia), GCHQ Problem Solving Group (U. K.), and the proposer.

The Least Common Multiple of *n* Consecutive Integers

10296 [1993, 291]. *Proposed by Paul Erdős, Hungarian Academy of Sciences, Budapest, Hungary.*

Let $f(n, a)$ denote the least common multiple of the n consecutive integers between a and $a+n-1$ inclusive.

(a) Show that for any positive integer k, there is a number $n_0(k)$ such that for each integer $n > n_0(k)$, there is an integer a with

$$f(n, a) > f(n, a+1) > \ldots > f(n, a+k).$$

(b)* Do there exist infinitely many integers n for which there is a pair of integers a, b with $a < b$ and $f(n, a) > f(n+1, b)$?

Solution to (a) by A. N. 't Woord, Eindhoven University of Technology, Eindhoven, The Netherlands. We prove that $n_0(k) = 2k$ will suffice. For $n > 2k$, put $a = pn! - k$, where p is an arbitrary positive integer. We need to establish the inequality $f(n, pn! - i) > f(n, pn! - i + 1)$ for $1 \le i \le k$. We use the notations $[.,.]$ and $(.,.)$ for the least common multiple and greatest common divisor, respectively.

Define $Q_i = f(n-1, pn! - i + 1)$. From the identity $(a, b)[a, b] = ab$, it follows that

$$f(n, pn! - i) = [pn! - i, Q_i] = \frac{(pn! - i)Q_i}{(pn! - i, Q_i)}$$

and

$$f(n, pn! - i + 1) = [Q_i, pn! - i + n] = \frac{Q_i(pn! - i + n)}{(Q_i, pn! - i + n)}.$$

Observe that

$$\big(a, f(n, a+1)\big) \;\Big|\; \prod_{i=1}^{n}(a, a+i) = \prod_{i=1}^{n}(a, i) \;\Big|\; n!,$$

and similarly

$$\big(f(n, a), a+n\big) \;\Big|\; \prod_{i=0}^{n-1}(a+i, a+n) = \prod_{i=0}^{n-1}(a+n, n-i) \;\Big|\; n!.$$

Hence $(pn! - i, Q_i) \mid (n-1)!$ and $(Q_i, pn! - i + n) \mid (n-1)!$. Also $pn! \mid Q_i$, so we have $(pn! - i, Q_i) = (pn! - i, (n-1)!) = i$ and $(Q_i, pn! - i + n) = (pn! - i + n, (n-1)!) = n - i$.

We may now write the inequality to be proved as $(n-i)(pn!-i) > i(pn!-i+n)$. This holds for $pn! > n^2$, for example, since $n > 2k \geq 2i$.

Solution to (b) by D. Eric Freeman (student), University of Michigan, Ann Arbor, MI, and Paul Lockhart, University of California, Santa Cruz, CA. There are infinitely many such n, since $f(p-2, p) > f(p-1, p+1)$ for all primes $p > 12$ that are congruent to 2 modulo 3. Using $[a, b, c, \ldots]$ for the least common multiple of $a, b, c, \ldots$, let $L = [p+1, \ldots, 2p-3]$. Then $f(p-2, p) = [p, L] = pL$ and $f(p-1, p+1) = [L, 2p-2, 2p-1]$. The integers $3(p-1)/2$ and $2(2p-1)/3$ belong to $\{p+1, p+2, \ldots, 2p-3\}$ and so must divide L. Hence $2p-2 \mid 4L$ and $2p-1 \mid 3L$. If follows that $[L, 2p-2, 2p-1] \leq 12L < pL$, and so $f(p-2, p) > f(p-1, p+1)$.

Editorial comment. The GCHQ Problem Solving Group provided a more elaborate construction which shows that for each positive integer k there exist infinitely many integers n for which $f(n, a) > f(n+k, b)$ for some $a < b$.

Other solutions to part (a) were given by M. Benito Muñoz & E. Fernández Moral (Spain), A. D. Melas (Greece), R. Stong, GCHQ Problem Solving Group (U. K.), and the proposer. Other solutions to part (b) were given by M. Benito Muñoz & E. Fernández Moral (Spain), D. Cass, T. Keller, A. D. Melas (Greece), A. Riese, R. Stong, A. N. 't Woord (The Netherlands), and the GCHQ Problem Solving Group (U. K.).

Another Stirling Property

10298 [1993, 400]. *Proposed by Donald E. Knuth, Stanford University, Stanford CA.*

Let $\left\{ {m+n \atop n} \right\}$ denote the number of ways to partition a set of $m+n$ elements into n nonempty subsets. Prove that

$$\frac{2^m 3^{\lfloor m/2 \rfloor} 4^{\lfloor m/3 \rfloor} 5^{\lfloor m/4 \rfloor} \cdots}{(n+1)(n+2)\ldots(n+m)} \left\{ {m+n \atop n} \right\}$$

is an integer.

Solution by Stephen M. Samuels, Purdue University, West Lafayette IN. The proof is in two steps. First we show that

$$\left\{ {m+n \atop n} \right\} = \sum (n+1)(n+2)\ldots(n+m) \binom{n}{a_1, \ldots, a_{m+1}} \prod_{r=2}^{m+1} \frac{1}{(r!)^{a_r}} \tag{1}$$

where the sum is taken over all choices of $m+1$ nonnegative integers $\{a_r\}$ such that

$$\sum_{r=1}^{m+1} a_r = n \quad \text{and} \quad \sum_{r=1}^{m+1} r a_r = n+m. \tag{2}$$

Then we show that, for all such choices of $\{a_r\}$,

$$\prod_{r=2}^{m+1} (r!)^{a_r} \quad \text{divides} \quad \prod_{r=2}^{m+1} r^{\lfloor m/(r-1) \rfloor}, \tag{3}$$

and hence the desired quantity is a sum of multiples of multinomial coefficients.

Condition (2) is clearly necessary and sufficient for the existence of a partition of $m+n$ elements into n sets in which a_r sets have size r. (Note that it is easily seen that no subset can have more than $m+1$ elements. This would also follow from equation (4), below, if the upper limit of the sum were left indeterminate.) For fixed $\{a_r\}$, the number of these partitions is the product of the number of ways of choosing, for each r, ra_r elements to be

in the sets of size r, and the number of ways of partitioning each such choice into a_r r-sets. This is

$$\binom{m+n}{a_1,\ldots,(m+1)a_{m+1}}\prod_{r=1}^{m+1}\frac{(ra_r)!}{(r!)^{a_r}}\cdot\frac{1}{a_r!},$$

which simplifies to the summand on the right side of (1).

We can rewrite $\prod_{r=2}^{m+1}(r!)^{a_r}$ as $\prod_{r=2}^{m+1}r^{a_r+\cdots+a_{m+1}}$, so showing $a_r+\cdots+a_{m+1}\le m/(r-1)$ yields (3). This follows from (2), because

$$(r-1)\sum_{j=r}^{m+1}a_j\le\sum_{j=1}^{m+1}(j-1)a_j\le m. \tag{4}$$

Editorial comment. The numbers $\left\{ {m \atop n} \right\}$ are called Stirling numbers of the second kind, and their explicit formula can also be derived from the well known generating function

$$\frac{(e^x-1)^n}{n!}=\sum_{m=n}^{\infty}\left\{ {m \atop n} \right\}\frac{x^m}{m!}.$$

Solved also by E. Alkan (student, Turkey), V. Božin (student, Yugoslavia), R. J. Chapman (U. K.), G, Evagelopoulos (Greece), R. Holzsager, I. Kastanas, O. P. Lossers (The Netherlands), A. D. Melas (Greece), and the proposer. One incorrect solution was received.

A Double Random Walk on an Odd Cycle

10299 [1993, 400]. *Proposed by José Luis Palacios, New Jersey Institute of Technology, Newark NJ.*

For an odd integer N, greater than 3, let G_N be the cyclic graph on N vertices. Consider the following random walk of *two* particles on G_N: at each time step, both particles independently move to one of the two adjacent vertices with probability 1/2.

If the particles initially occupy adjacent vertices, what is the expected number of jumps until the particles meet?

Solution by Daniel L. Stock, Rocky River, OH. The expected number of time steps is $(N^2-1)/2$. Let $k=(N-1)/2$, and for $0\le i\le k$ let e_i denote the expected number of steps until the particles meet if the path between them of even length has length $2i$. Then $e_0=0, e_1=1+.5e_1+.25e_2, e_k=1+.25e_{k-1}+.75e_k$, and $e_i=1+.25e_{i-1}+.5e_i+.25e_{i+1}$ for $2\le i\le k-1$. This yields k linear equations in k unknowns, or the matrix equation

$$\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{k-1} \\ e_k \end{pmatrix}=\begin{pmatrix} 4 \\ 4 \\ \vdots \\ 4 \\ 4 \end{pmatrix}.$$

Replacing each row by its sum with all the rows below it yields

$$\begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & & \\ & -1 & 1 & \\ & & -1 & 1 \end{pmatrix}\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{k-1} \\ e_k \end{pmatrix}=\begin{pmatrix} 4k \\ 4k-4 \\ \vdots \\ 8 \\ 4 \end{pmatrix}.$$

From this, the solution is $e_i=\sum_{j=0}^{i-1}4(k-j)=4ki-4\binom{i}{2}$. Setting $i=k$, we obtain $e_k=2k(k+1)=(N^2-1)/2$.

Editorial comment. Ellen Hertz observed more generally that if the initial path of odd length between the particles has length j, then we can set $i = (N - j)/2$ in the solution above and obtain $(N^2 - j^2)/2$ as the expected time to collision. James Kuttler observed that the first matrix above is well known in finite-difference approximations and has an inverse in which the entry in position i, j is $\min\{i, j\}$. Several solvers reduced the problem to known results about the "gambler's ruin" problem.

Solved by 41 other readers and the proposer. Three incorrect solutions were received.

Collaborating editors: *David F. Appleyard, Paul T. Bateman, Duane M. Broline, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, and William E. Watkins.*

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

Hypatia of Alexandria. By Maria Dzielska. Volume 8 of *Revealing Antiquity*, Harvard University Press, 1995, viii + 157, $29.95.

Reviewed by **Michael A. B. Deakin**

We have waited over two centuries since the last book-length biography of Hypatia of Alexandria was published in English. Indeed, although there is much extant material for the sufficiently patient contemporary researcher to unearth, almost all of it is hard to come by, written in foreign languages or else utterly unreliable. So we welcome the appearance (finally) of a book, from a scholarly author and issued under a reassuring imprint, on one of the very first women mathematicians known to history.

Hypatia lived in the late 4th and early 5th centuries of our era. She was domiciled throughout her life in Alexandria which, although geographically in Egypt and at the time part of the Roman Empire, was culturally Greek. The traditions she had inherited began with Euclid, who could be described in modern terms as the first 'professor' of mathematics in the university of Alexandria (better known by its transliterated Greek name, the 'Museum'). The Museum endured from Euclid's time (about 300 BC) for some seven centuries. Its last attested member, perhaps its last 'president', was a mathematician: Theon of Alexandria.

As these things are judged, Theon was a minor talent, but nonetheless an important one; for he conserved and promulgated the primary texts of Euclid and of Ptolemy and thus is responsible in large measure for their present day accessibility. Theon was Hypatia's father, and she was very much his daughter. Theon edited and produced 'Commentaries' on the great Alexandrian mathematical classics: the works of Euclid and Ptolemy. Hypatia extended this 'great books' programme to the later and more difficult expositions of Apollonius and Diophantus. We know this because an ancient encyclopedia (the so-called *Suda Lexicon*) tells us so; it also tells us (in a rather garbled passage) that Hypatia had something to do with 'astronomical tables'.

So here is Hypatia, author of 'cribs' to Apollonius and to Diophantus, and some (whatever) exercise on astronomical tables: a mathematician within the meaning of the act. Not necessarily a great mathematician, but clearly one of the fold, and in fact the pre-eminent mathematician of her day. We have no evidence that Hypatia ever wrote on any subject beyond mathematics and astronomy. (It is thus misleading to say, as Dzielska does, 'No titles of her philosophical works are extant' (p. 54) —there may very well have been none. From her published work, Hypatia was first and foremost a mathematician.)

But it is equally true that Hypatia was not *only* a mathematician. She also took up the study of other (non-mathematical) branches of philosophy, the *Suda* tells us. The word 'philosophy' could be used in a generic sense to mean 'learning in general' or it could be applied in the narrower sense we employ today. The non-mathematical philosophy Hypatia espoused was of this latter sort. This philosophy was neoplatonist, by which is meant that it was of a religious character. We can know something of the philosophical system she expounded from an examination of the writings of her pupil Synesius, many of which survive.

This aspect of her life in fact ultimately led to her death. Alexandria was at that time desperately faction-ridden, with Christians, both orthodox and heretical, various groups of 'pagans' (among whom the neoplatonists were included), Jews and others; and all of them at one another's throats. Against this background, and in the course of a conflict between the ecclesiastical and the civil authorities, a mob of Christian fanatics, charging her with the practice of witchcraft, brutally murdered her.

These events have made her a symbol and a powerful one for intellectuals, feminists and, of course, anti-clericals. Much of the writing that has invoked her name is essentially fiction, or generically legend. Dzielska devotes quite a large part of her biography to an account of the various forms this legend has taken, before proceeding to a historically informed analysis of her life and thought. It is this later part of the book that holds most interest.

There is a lot to be grateful for in this well-presented and readable account. Many of the conclusions challenge accepted, or at least widely promulgated, wisdom, and furthermore they make a lot of sense. For example, the date of Hypatia's birth is correctly placed well before the 370 AD usually quoted. (I was a little surprised, however, that there was *no discussion at all* of the date of Hypatia's death. This is normally quoted as 415 AD, a date Dzielska adopts uncritically. The best and most detailed discussions, however, opt for 416.)

There is a brave and well-founded attempt to reconstruct aspects of her philosophical (i.e. neoplatonist) system of beliefs. (Unlike, say, Christianity, neoplatonism made little attempt to enforce uniformity of doctrine, and so the description of a person as a neoplatonist says very little about the actual beliefs held.)

Central to all neoplatonist philosophy however is the activity of abstraction. We form general notions (e.g. of the number 2) by abstracting from observed instances (in this case pairs). These abstractions become ideas, or ideals, and these were seen by neoplatonists as constituting a deeper and more fundamental reality than the world of appearance which for them was merely a shadow, or projection, of that underlying reality.

A further process of abstraction then leads us from the ideas to a yet deeper, even more fundamental, level of reality. This is the idea of ideas, the One. Properties attributed to the One make it a sublime ineffable essence, not at all unlike the God of monotheistic tradition. This is the basis of the religious aspect in neoplatonism.

The mathematical example is surely the best that can be given. Almost certainly, most working mathematicians espouse a vaguely platonic philosophy toward the objects of their contemplation. It informs the very language of mathematical discourse. We speak of 'the properties of the circle' as if 'the circle' were something quite real, to which it is quite natural to attribute properties; we discuss 'the discovery of Pythagoras' theorem' as if, prior to this, Pythagoras' theorem were so to speak sitting out there like America waiting to be discovered.

Thus for many neoplatonists mathematics held a special place as a route to the One, which was the ultimate goal of all human endeavour. It would seem natural to think Hypatia espoused such a view, and indeed there is historical evidence to support this hypothesis. Certainly such a view is very much part of the thinking of Synesius of Cyrene, Hypatia's best-known pupil. Hypatia is the only teacher Synesius ever mentions and he almost palpably worshipped her; he sent her copies of two of his works, in effect asking her to referee them. We may assume that they met with her approval, for they are still extant today.

If therefore we are to reconstruct Hypatia's philosophy, the best place to start is with that of Synesius. This obvious methodology is that adopted by Dzielska, and to my mind she succeeds admirably. Many other such discussions, by contrast, are predicated on the wildest or flimsiest of hypotheses and almost every conceivable suggestion that could be advanced has been. Thus much of Dzielska's account is necessarily devoted to the refutation of error.

But then Dzielska herself falls into a similar trap the moment she begins the discussion of Hypatia's mathematics. She does not, like almost all earlier biographers, shirk the issue. Its importance is recognised. The trouble is that she does not come equipped with the necessary competences to deal with mathematical or scientific matters.

We know that Hypatia wrote on arithmetic (Diophantus), geometry (Apollonius) and on astronomy (the 'table'). We also know that she helped Synesius to design an astrolabe (which he had crafted by the best silversmith he could find and which he presented to one Paeonius, an official with whom he wished to curry favour—the covering letter survives). Also still with us is his request to Hypatia to have a hydrometer constructed for him (for medical purposes during his final illness).

Even long before this time, arithmetic and geometry had been thoroughly 'platonized'. Euclid's *Elements* deals with a platonic world of lines, circles, prime numbers and other such, with virtually no reference to the concrete instantiations that pertain to the everyday world. This is not so with astronomy, however. It has always, then as now, retained its material referents: the sun, moon, stars and planets. These bodies were seen as material entities, composed of a quintessence, a fifth element more noble than earth, air, fire and water, but material nonetheless.

Thus the study of astronomy, though worthy and a path to higher things, was *not* viewed as being superior to geometry, as Dzielska (p. 54) asserts Hypatia to have held. It's the other way round, and Dzielska has misunderstood the situation. She goes on to quote Synesius' covering letter to Paeonius. But here, in more detail, is what Synesius said: 'Astronomy itself is a venerable science, and might become a stepping stone to something more august, a science which I think is a convenient passage to mystic theology.... It proceeds to its demonstrations in no uncertain way, for it uses as its servants geometry and arithmetic, which it would not be improper to call a fixed standard of truth'. (The final sentence cites Ptolemy's *Almagest*.) Although the master-servant relation is invoked, it is nevertheless quite clear which branches of mathematics embrace the higher truth.

Although Theon, like his daughter, is described as a 'philosopher' in some of the sources, the term is here used generically. Theon was a specialist mathematician. Many of the non-mathematical works that Dzielska attributes to him are more usually ascribed to other *Theons*. She also relies too easily on a brief sentence in the account of the 6th-century historian John Malalas. And so we come away with the wrong picture of Theon. Moreover the sources (e.g. the *Suda*)

are quite explicit: what *distinguished* Hypatia from Theon was her unique synthesis of the mathematical with the philosophic.

But here we are offered *no* real discussion of Hypatia's mathematics at all; although brief reference is made to the work of Toomer and of Knorr, no acquaintance with that work is evident. Knorr has done much to reconstruct Hypatia's mathematics, although this work has been criticized by Cameron. Toomer gives us a much better picture of Theon. Much of Dzielska's work derives from Cameron's, but a lot more attention should have been paid to Knorr's and to Toomer's.

Instead we find things like the silly sentence (p. 71): 'Apollonius' work *The Conic Sections* was in trigonometry; Perl has attempted to reconstruct Hypatia's commentary on it'. The first part of this is at (very) best misleading, the second downright wrong. Teri Perl's *Math Equals* is an imaginative high school textbook that integrates didactic material with brief derivative biographies of famous women mathematicians (that on Hypatia being best forgotten). It has nothing to do with the 'reconstruction' of Hypatia's lost work!

The aim of science, in Hypatia's day as in ours, was in large measure the quest to predict the future. The exact sciences succeed to the extent that they are predictive. Astronomy was the first such success, and it did spectacularly well; by Ptolemy's day the positions of the planets could be predicted with very high accuracy and even eclipses could be foretold. Attempts to extend this program into other areas, areas of more immediate human concern, however led to astrology and to other techniques of divination. As part of all this activity, people set up as 'mathematicians' who were actually, in our terms, astrologers or numerologists.

Both the Christian church and the secular authorities tried to put a stop to this. The emperor Constantius decreed that 'no-one may consult a soothsayer or a mathematician'; St. Augustine of Hippo (a contemporary of Hypatia's) did consult 'mathematicians', squaring this initially with his conscience via a neat bit of casuistry, but when he discovered them to be charlatans, he decided that mathematics was immoral after all!

Now such 'mathematics' is very different from writing commentaries on Apollonius or Diophantus, and it is this latter variety that Hypatia and her circle practised (although the charge of witchcraft against her arose from confusion between mathematics and 'mathematics'). Dzielska for the most part recognizes this, but every so often she backs away from this position and so she has Theon indulging in magic and Synesius resorting to hydromancy.

We do know that Ptolemy indulged in astrology as well as reputable astronomy; he even wrote a book, the *Tetrabiblos*, on the subject. Theon wrote commentaries on Ptolemy's *Almagest* and others of his astronomical works, but did not comment on or edit the *Tetrabiblos*. It would seem that his alleged interest in astrology is highly questionable. But Dzielska's picture of Theon is coloured by her uncritical approach to marginal and dubious sources.

Dzielska's evidence for 'hydromancy' on Synesius' part is little more than an interpretation of the use of the hydrometer. Passed over completely are the more plausible suggestions that it was used to test the quality of Synesius' drinking water, or was involved in the production of an alcohol-based medicine by brewing or distillation, or (as I believe) that it was intended as a urinometer. To have Synesius resort to 'hydromancy' is not only fanciful, but also at variance with her other assessments of Synesius' character and beliefs.

The failure of Dzielska's book to some to terms with mathematical and scientific matters is disappointing. All in all, however, the book has much good sense in it and is a welcome addition to the literature.

Department of Mathematics
Monash University
Clayton, Victoria 3168
AUSTRALIA

Notes on Set Theory. By Yiannis N. Moschovakis. Springer-Verlag, 1994, vii + 272, $39.00

Reviewed by **Joan E. Hart and Kenneth Kunen**

How much set theory do you need to know? Should you read this book? To help you answer these questions, we partition, rather arbitrarily, basic set theory into "elementary," "intermediate," and "advanced," and we touch on the relevance of set theory to philosophy and computer science. Then we comment on what Moschovakis includes, and we conclude with some additional remarks on the exposition.

Elementary: Children learn rather quickly how to count the cookies in a jar:

$$0, 1, 2, 3, \dots .$$

By high school, students know that they could call the jar a "set," and they know some basic facts about unions and intersections, and how these relate to the sizes (cardinalities) of sets. Related to this are some facts from discrete math, such as what a function is, and what it means for a relation on a set to be a total ordering, reflexive, etc. This much set theory will take students through the basic college courses in calculus and abstract algebra, which, after all, cover material primarily discovered before Cantor. It will also take them through elementary courses in computer programming, where they learn how to represent functions, relations, and finite sets in Basic or C.

Intermediate: Eventually, it becomes important to know something about post-Cantorian set theory. Every math major should learn that the set of reals, $\mathbb{R}$, cannot be covered by a countable sequence of points (Cantor), or even by a countable sequence of nowhere dense sets (Baire). Undergraduates often learn, in addition, how to recite Zorn's lemma and how to use it to prove, for example, that every vector space has a basis. This material is often covered in introductory courses in real analysis, topology, or algebra, and is all the set theory that most research mathematicians ever need to know.

Advanced: In some areas of mathematics, one needs to know about well-orderings and ordinals, and how to count the cookies in an infinite jar:

$$0, 1, 2, 3, \dots, \omega, \omega + 1, \dots, \omega + \omega, \dots, \omega_1, \dots, \omega_2, \dots .$$

If the jar is $\mathbb{R}$, it has size ω_α (or $\aleph_\alpha$) for some $\alpha > 0$; the Continuum Hypothesis says that $\alpha = 1$. These topics are covered in detail in an undergraduate course in set theory. Unfortunately, most math majors never take such a course. As a result, even graduate level texts avoid using these notions, although a number of topics would be made somewhat less obscure by a knowledge of ordinals. For example, Zorn's lemma seems rather mystical unless you understand how to prove it using ordinals; and, if you understand that proof, you no longer need Zorn's lemma; you just pick a basis for an infinite dimensional vector space inductively, exactly the same way you pick a basis for a finite dimensional one. Or, every book on measure theory states that the Borel sets form the least σ-algebra containing the open sets, but constructing the Borel sets inductively in ω_1 steps (open sets, G_δ sets, G_{δ_σ} sets,...) gives you a much clearer picture of what they are than does producing them by intersecting a family of σ-algebras.

Philosophy: Mathematicians should have some understanding of the foundational underpinnings of their art. Although the axiomatic method in geometry goes back to Euclid, the modern view is that all of mathematics can be developed within the unified framework of axiomatic set theory. One does not introduce a whole new collection of axioms for each mathematics course. One starts with something like the Zermelo-Fraenkel axioms (ZF), and proceeds both to develop general abstract facts about sets and functions, as well as to define important specific sets, such as the natural numbers, the rational numbers, and the real numbers, and then the Euclidean plane ($\mathbb{R} \times \mathbb{R}$). So, geometry, like everything else, is a branch of set theory. This axiomatization also reveals the role of the Axiom of Choice (AC). Whether or not one admits AC as a basic principle, one should understand which results from elementary mathematics require AC.

Computer Science: Of course, anything that is *done* on the computer is finite, and can be understood using just elementary set theory, but more advanced methods come in when trying to understand the theory behind what the computer is doing. Thus, books on denotational semantics for programming languages ([**5**]) use the kind of set-theoretic techniques usually associated with general topology. Ordinals crop up in books on logic programming semantics ([**8**]) and implementations of constructive mathematics ([**2**]).

This book starts off in the beginning of the intermediate level, which is ideal for an undergraduate text. The first chapter quickly reviews elementary facts about sets and functions, primarily to establish the notation to be used. The second chapter explains Cantor's basic ideas, covering countable and uncountable sets, and Cantor's diagonal argument. The explanatory remarks and accompanying figures should be very helpful for readers who haven't seen these things before.

The third chapter points out that naive manipulation with sets can lead to contradictions, such as Russell's paradox with $\{x: x \notin x\}$, hence the need for some axiomatic framework. Moschovakis explains what is meant in general by an axiomatic system, and then describes Zermelo's axioms for set theory.

By the end of Chapter 5, the reader sees how, based on Zermelo's axioms, one can develop elementary discrete math (sets, functions, relations), as well as the natural numbers, $\mathbb{N}$. In particular, Moschovakis stresses that the existence of $\mathbb{N}$, along with the basic facts about $\mathbb{N}$ (such as induction and recursion), are all theorems within the framework of axiomatic set theory. At this point, one has the machinery to go on and develop the rational numbers and the real numbers, but

this requires some knowledge of algebra, and is put off until Appendix A, where it is done in detail, using both Dedekind cuts and Cauchy sequences.

Of course, every book on set theory would have to tell you this much; the only serious design decision is whether to develop $\mathbb{N}$ first and then cover the transfinite ordinals later, as a more advanced topic, or whether to plunge right in to the general theory of ordinals, obtaining $\mathbb{N}$ as the set of the first ω ordinals. Moschovakis chooses the first option, which is slightly redundant, but which makes the material more accessible for undergraduates approaching this abstract subject for the first time.

Chapter 6 is an important departure from tradition. It is centered around the notion of an *inductive poset*; this is a partially ordered set, P, such that every chain (totally ordered subset) has a least upper bound. The basic result here is that every monotonic mapping on an inductive poset has a least fixed point. It is rather unusual to see this in an undergraduate mathematics text, since the most well-known applications are in computer science. The exercises give a hint of how this is applied in programming language semantics. The text gives the following very concrete application: In most programming languages, you can define a function f *recursively*, defining $f(x)$ by any expression which involves f itself. In general, such a computation might fail to terminate for some (maybe all) values of x; the fixed point theorem shows that every such definition uniquely determines f as a *partial* function. If the input and output to f are natural numbers, then the relevant P is the set of all partial functions on $\mathbb{N}$, ordered by subset (if we identify each partial function with its graph).

Besides its interest in computer science, fixed points and inductive posets are a nice way of introducing the mathematics topics which follow. Chapter 7 introduces well-orderings. Fixed point theory yields a motivation for the study of well-orderings, since you actually need well-orderings to prove the fixed-point theorem (Chapter 6 only proves a weakened version of the theorem). Then, Chapter 8 introduces AC. It proves the standard equivalents to AC (such as the well-ordering principle and Zorn's lemma), but the proofs, using the poset and fixed point terminology, are a good deal more elegant than the standard ones.

Chapter 9 uses AC to develop some further material, such as the theory of cofinalities, and König's Theorem. For example, although Gödel and Cohen tell us that the continuum, $2^{\aleph_0}$, could be $\aleph_1$ or $\aleph_5$ or $\aleph_{\omega+1}$, it cannot be $\aleph_\omega$ or $\aleph_{\omega+\omega}$ by König's Theorem.

Chapter 10 covers basic descriptive set theory. It is another important departure from tradition to do this in an undergraduate text. This material is definitely learnable on an undergraduate level, and is something which every mathematician should know, but often doesn't. A typical result here is that the Continuum Hypothesis is simply a *theorem* for Borel sets. That is, every Borel subset of $\mathbb{R}$ is either countable or contains a perfect subset, and hence has size $2^{\aleph_0}$. Chapter 10 also presents the construction of a Bernstein set (an uncountable $X \subset \mathbb{R}$ such that neither X nor $\mathbb{R} \setminus X$ contains a perfect subset); this is not descriptive set theory, but is a nice application of well-ordering and transfinite induction to elementary real analysis.

Chapter 11 introduces the Axioms of Replacement and Foundation, which are part of ZF, but not part of Zermelo's original axioms. Chapter 12, finally, tells you about von Neumann ordinals and cardinals. Here, you learn what the ordinals really are; for example, $0 = \varnothing$; $3 = \{0, 1, 2\}$, and $\omega = \mathbb{N}$. Appendix B is an introduction to models of set theory and consistency proofs; in particular, it

describes models where Foundation is true, and other models where Foundation is false.

In general, the text's conversational style makes it easy to read, and its content is instructive. Moschovakis often introduces results with remarks on their importance, then guides the reader through the proofs, helping students not only to follow the steps, but also to learn common proof techniques. For example, in Chapter 5, Moschovakis talks the reader through a proof of the Recursion Theorem, showing the reader how to obtain a function from the finite partial functions which approximate it; this material forms a good introduction to the more sophisticated kinds of recursion covered in Chapters 6 and 7. In addition, he is careful to point out key ingredients to various proofs. For example, after proving the uncountability of the set of real numbers, he stresses the role the completeness property of the reals plays in the proof, noting "the rest of Cantor's construction relies solely on arithmetical properties of numbers which are also true of rationals."

In relating some of the history of set theory, Moschovakis gives the reader insight into the roots of the subject. He points out that Cantor's approach was rather vague and intuitive, so that mathematicians of the day were naturally suspicious of his methods. His discussion helps the student understand how the paradoxes led to Zermelo's formulation in 1908 of a precise set of axioms. These axioms spelled out exactly what is being assumed, and seemed to be free of contradictions. To be fair to Cantor though, Moschovakis should have noted how close Russell's Paradox is to what Cantor already knew. Chapter 2 presents Cantor's diagonal argument that there is no map π from a set A onto its power set, $\mathscr{P}(A)$: the assumption that $B = \{x \in A\colon x \notin \pi(x)\}$ is in the range of π leads to a contradiction. Cantor knew that there was a problem if A is the universal set, V, since then $\mathscr{P}(V)$ is a subset of V, not bigger than V. More succinctly, if π is the identity map and $A = V$, we get an outright contradiction from $B = \{x \in V\colon x \notin \pi(x)\} = \{x\colon x \notin x\}$. Of course, Cantor left it to Russell to put the paradox this succinctly, and then to popularize it. But, it is misleading to state in Chapter 3 that the paradoxes before Russell were "technical and affected only the most advanced parts of Cantor's theory." It would be more accurate to say that Cantor's methods were informal and intuitive, and that he just intuitively avoided what he called "inkonsistenten Mengen" (see **[1]** for further discussion).

The descriptive set theory in Chapter 10 focuses on *Baire space*, $\mathscr{N} = \mathbb{N}^{\mathbb{N}}$, a countable product of countable discrete spaces. It will be a little difficult for most readers to see that the results also apply to more familiar spaces, such as $\mathbb{R}$. Moschovakis does mention that "there is such a tight connection between $\mathscr{N}, \mathscr{C}$ [the Cantor set], and $\mathbb{R}$ that practically every interesting property of one of these spaces translates immediately to a related, interesting property of the others." But, the details are a bit patchy, and must be ferreted out of the problems (x10.11, x10.12) and Appendix A. Neither in the problems nor in Appendix A does he point out that $\mathscr{N}$ is homeomorphic to the space of irrational numbers, although the classical proof of this result maps $\mathscr{N}$ to the irrationals by a simple use of continued fractions, a topic of numerous *Monthly* articles and one definitely accessible at the undergraduate level.

Moschovakis does not introduce ordinal notation until the last chapter (12), which means that we can't really count our cookies in the standard way until the end of the book, where we finally see $\omega, \omega + 1, \omega + 2, \ldots$. As he says in the Preface, most courses will not get this far. This is unfortunate, since this method of counting is frequently used when a transfinite sequence is listed. Its roots go back to Cantor's original theory; even in his paper of 1880, Cantor employed such

"infinite symbols" to advance the theory of derived sets, three years before his *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* presented the transfinite numbers as "the simplest, most appropriate and natural extension [of the concept of number"] ([3]). There is a formal justification for Moschovakis' order of presentation: to develop the modern (von Neumann) theory of ordinals (as opposed to Cantor's intuitive presentation), one needs the Replacement Axiom, which is not introduced until Chapter 11, and it is of some formal interest (to specialists) to see how much set theory can be developed without Replacement. However, in a number of places in the text, especially in Chapters 7 and 9, the explanations would have been much simpler and more natural if ordinal notation were available. For example, without ordinals, the discussion of cofinalities in Chapter 9 is a bit awkward, and the motivation for studying König's Theorem is a bit obscure; since the notation $\aleph_\alpha$ could not be defined yet, he was actually not able to state simply that $2^{\aleph_0} \neq \aleph_{\omega+\omega}$, as we did above.

Finally, the book is only 272 pages long, and cannot cover everything. Overall, the author has made an excellent choice of what to include, and he says just enough about omitted topics to whet the reader's appetite for more. So, the reader may be disappointed to find no references to the literature beyond the two historical sources cited in the Preface. For example, the author mentions logic, including the fact that the Gödel incompleteness theorems apply to systems such as *ZF*; why not suggest an undergraduate logic text (e.g. [4]) where the reader might pursue the subject further? Or, the book tells us that by results of Gödel and Cohen, *CH* is true in some models of *ZFC* and false in others, but there are no references to texts such as [6] or [7], where these models are constructed. Given the relevance of Chapter 6 to programming language semantics, it seems strange not to refer the reader to a basic text ([5]) on the subject. Stranger still, many of the missing references were written by the author himself, who is one of the leading contributors to many of the topics highlighted in the book, such as fixed point theory ([9]), descriptive set theory ([10]), and applications of logic to computer science ([11]).

REFERENCES

1. A. A. Fraenkel, Y. Bar-Hillel, and A. Lévy, *Foundations of Set Theory*, Second Edition, North-Holland, 1984.
2. R. S. Boyer and J. S. Moore, *A Computational Logic Handbook*, Academic Press, 1988.
3. J. W. Dauben, *Georg Cantor, His Mathematics and Philosophy of the Infinite*, Princeton University Press, 1979.
4. H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, 1972.
5. C. A. Gunter, *Semantics of Programming Languages*, The MIT Press, 1992.
6. T. Jech, *Set Theory*, Academic Press, Inc., 1978.
7. K. Kunen, *Set Theory, An Introduction to Independence Proofs*, North-Holland, 1980.
8. J. W. Lloyd, *Foundations of Logic Programming*, Second Edition, Springer-Verlag, 1987.
9. Y. N. Moschovakis, *Elementary Induction on Abstract Structures*, North-Holland, 1974.
10. Y. N. Moschovakis, *Descriptive Set Theory*, North-Holland, 1980.
11. Y. N. Moschovakis, editor *Logic from Computer Science*, Springer-Verlag, 1992.

Department of Mathematics
University of Wisconsin
480 Lincoln Drive
Madison, WI 53706-1388
jhart@math.wisc.edu
kunen@cs. wisc.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee and Paul Zorn**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly.*

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *Conversations on Mind, Matter, and Mathematics.* Jean-Pierre Changeux, Alain Connes. Ed. & Transl: M.B. DeBevoise. Princeton Univ Pr, 1995, xii + 260 pp, $24.95. [ISBN 0-691-08759-8] A fascinating colloquy between two creative individuals on the nature of knowledge and reality, using mathematics (Connes) and neuroscience (Changeux) as the playing field for philosophical jousting. Evolution, intelligence, language, and logic compete under the traditions of Platonism and Darwinism in a search for the material sources of mathematical thought. LAS

General, S*, L.** *New Mathematical Diversions From Scientific American, Revised Edition.* Martin Gardner. Spectrum Ser. MAA, 1995, 268 pp, $19.95 (P). [ISBN 0-88385-517-8] Reprint of 1983 edition, with new postscript and bibliography (University of Chicago Press edition, TR, January 1985). LCL

Reference, P, L*. *Encyclopaedia of Mathematics, Volumes 6–10.* Ed: M. Hazewinkel. Kluwer Academic, $199 each. *Volume 6: Lobachevskiĭ Criterion (for Convergence)—Optional Sigma-Algebra,* 1990, ix + 546 pp, [ISBN 1-55608-005-0]; *Volume 7: Orbit-Rayleigh Equation,* 1991, ix + 506 pp, [ISBN 1-55608-006-9]; *Volume 8: Reaction-Diffusion Equation—Stirling Interpolation Formula,* 1992, ix + 540 pp, [ISBN 1-55608-007-7]; *Volume 9: Stochastic Approximation—Zygmund Class of Functions,* 1993, ix + 536 pp, [ISBN 1-55608-008-5]; *Volume 10: Subject Index—Author Index,* 1994, xiii + 732 pp. [ISBN 1-55608-009-3]

Mathematics Appreciation, T, L. *Five Golden Rules: Great Theories of 20th-Century Mathematics—and Why They Matter.* John L. Casti. Wiley, 1995, xii + 225 pp, $24.95 (P). [ISBN 0-471-00261-5] Discusses five theorems (minimax, Brouwer's fixed-point, theorems of Morse and Halting, and the simplex method), their solutions and their significance. Intended for non-mathematicians; appropriate for "liberal arts" mathematics. LC

Education, S(13–18). *Prospects for School Mathematics.* Ed: Iris M. Carl. NCTM, 1995, ix + 340 pp, $20. [ISBN 0-87353-418-2] 21 essays on K–16 students, teaching, content, and context. Views of mathematics education aligned with the NCTM *Standards* documents. Educators, researchers, and policy analysts offer perspectives on change, citing both ideas and challenges for achieving excellence with equity. Highlights include short history of 20th-century reform, broad bibliographies. MW

Education, P. *Constructing Mathematical Knowledge: Epistemology and Mathematics Education.* Ed: Paul Ernest. Stud. in Math. Educ. Ser., V. 4. Falmer Pr, 1994, xviii + 282 pp, $65. [ISBN 0-7507-0354-7] Multiple perspectives on learning mathematics; in 19 chapters, each with different author(s). Includes constructivist and social theories, the influence of the computer, and the role of language in learning mathematics. Explores inquiry in mathematics education, and relates teaching and learning of mathematics with its historical development. Along with companion Volume 3 in this series, *Mathematics, Education and Philosophy: An International Perspec-*

tive, it aims to "set a research agenda for the philosophy of mathematics education." MW

Education, S(13–14). *Investigating Mathematics.* Howard Zeiderman. CZM Pr, 1994, viii + 150 pp, $22 (P). [ISBN 1-878461-20-6] Designed as a supplementary text for senior high students, this volume could profitably supplement instruction of future elementary and secondary teachers or mathematics appreciation students. 16 units cover such topics as definitions and metaphors, mathematical thinking and ability, nature of mathematical truth. Each unit contains an orientation, a short historical text, and questions for discussion. Through discussion, students see mathematics as a "deeply human activity," requiring decisions, initiative, and "self-examination of presuppositions and intellectual biases." MW

History, S(15–16), L. *Quantification and the Quest for Medical Certainty.* J. Rosser Matthews. Princeton Univ Pr, 1995, x + 195 pp, $39.50. [ISBN 0-691-03794-9] History of debates surrounding the development of medical statistics and the clinical trial. LB

Logic, P. *Non-Classical Logics and Their Applications to Fuzzy Subsets: A Handbook of the Mathematical Foundations of Fuzzy Set Theory.* Eds: Ulrich Höhle, Eric Peter Klement. Theory & Decision Lib., Ser. B, Math. & Stat. Methods, V. 32. Kluwer Academic, 1995, viii + 390 pp, $105. [ISBN 0-7923-3194-X] Proceedings of the 14th Linz Seminar on Fuzzy Systems (September 1992).

Logic, P, L. *Kurt Gödel: Collected Works, Volume III: Unpublished Essays and Lectures.* Eds: Solomon Feferman, *et al.* Oxford Univ Pr, 1995, xix + 532 pp, $59.95. [ISBN 0-19-507255-3]

Foundations, P, L. *The Universal Turing Machine: A Half-Century Survey, Second Edition.* Ed: Rolf Herken. Springer-Verlag, 1995, xvi + 611 pp, $49 (P). [ISBN 3-211-82637-8] Corrected republication of the 1988 Oxford University Press edition (TR, April 1989).

Number Theory, P. *Infinity and the Mind: The Science and Philosophy of the Infinite.* Rudy Rucker. Princeton Univ Pr, 1995, xii + 342 pp, $12.95 (P). [ISBN 0-691-00172-3] Paperback republication of 1982 Birkhäuser original (TR, November 1982; Extended Review, February 1984).

Number Theory, P. *Proceedings of the St. Petersburg Mathematical Society, Volume III.* Ed: O.A. Ladyzhenskaya. Transl. Ser. 2, V. 166. AMS, 1995, ix + 267 pp, $92. [ISBN 0-8218-0387-5] 12 papers by students of D.K. Faddeev, and by their students.

Linear Algebra, T(14: 1). *Linear Algebra: An Introduction.* Richard Bronson. Academic Pr, 1995, xi + 504 pp, $39.95 (P). [ISBN 0-12-135245-5] A careful, logically ordered presentation of traditional topics (matrices, vector spaces, linear transformations, eigenvalues and eigenvectors, inner products, Jordan canonical forms) without reference to computing. JNC

Group Theory, P. *Groups '93 Galway/St Andrews: Galway 1993.* Eds: C.M. Campbell, *et al.* Cambridge Univ Pr, 1995. *Volume 1*, London Math. Soc. Lect. Note Ser., V. 211, xii + 304 pp, $39.95 (P), [ISBN 0-521-47749-2]; *Volume 2*, London Math. Soc. Lect. Note Ser., V. 212, xii + 304 pp, $39.95 (P). [ISBN 0-521-47750-6] Selected papers from the August 1993 international conference.

Algebra, P. *Zero-Dimensional Commutative Rings.* Eds: David F. Anderson, David E. Dobbs. Lect. Notes in Pure & Appl. Math., V. 171. Marcel Dekker, 1995, xi + 378 pp, $135 (P). [ISBN 0-8247-9605-5] Proceedings of the 1994 John H. Barrett Memorial Lectures and Conference on Commutative Ring Theory at the University of Tennessee, Knoxville.

Algebra, P. *Cohomological Induction and Unitary Representations.* Anthony W. Knapp, David A. Vogan, Jr. Math. Ser., V. 45. Princeton Univ Pr, 1995, xvii + 948 pp, $69.50. [ISBN 0-691-03756-6]

Algebra, P. *Second International Conference on Algebra: Dedicated to the Memory of A.I. Shirshov.* Eds: L.A. Bokut', A.I. Kostrikin, S.S. Kutateladze. Contemp. Math., V. 184. AMS, 1995, xi + 449 pp, $65 (P). [ISBN 0-8218-0295-X] Proceedings of the 1991 conference in Barnaul, Russia.

Algebra, T(14–15: 1), L. *Boolean Algebra and Its Applications.* J. Eldon Whitesitt. Dover, 1995, x + 182 pp, $7.95 (P). [ISBN 0-486-68483-0] Presents Boolean algebra as an abstract algebraic system. Three chapters on use of Boolean algebra in designing switching, relay, and computer circuits. Introduces algebra of sets, symbolic logic, and discrete probability. Republication of 1961 edition. KES

Algebra, P. *Third Siberian School: Algebra and Analysis.* Eds: L.A. Bokut', M. Hazewinkel, Yu. G. Reshetnyak. Transl. Ser. 2, V. 163. AMS, 1995, x + 188 pp, $95. [ISBN 0-8218-0286-0] 16 lectures presented in 1989.

Algebra, T(17–18: 1), S, P.** *Polynomial Invariants of Finite Groups.* Larry Smith.

Res. Notes in Math., V. 6. AK Peters, 1995, xvi + 360 pp, $59.95. [ISBN 1-56881-053-9] A nicely written treasure trove of classical and modern results, using (and motivating!) much modern mathematics—algebraic geometry, commutative algebra and ideal theory, homological algebra, algebraic combinatorics. Many useful worked examples. RM

Complex Analysis, T(17–18: 1, 2), S, P, L. *Functions of One Complex Variable II.* John B. Conway. Grad. Texts in Math., V. 159. Springer-Verlag, 1995, xvi + 394 pp, $49. [ISBN 0-387-94460-5] A brisk but clear and readable sequel to author's introductory graduate text. First 5 chapters continue first volume's thread, culminating in de Branges's proof of the Bieberbach conjecture. Second half, emphasizing harmonic and subharmonic functions, assumes more analysis and functional analysis. Each section ends with a few exercises. PZ

Complex Analysis, T(18), S, P*, L. *Topics in Hardy Classes and Univalent Functions.* Marvin Rosenblum, James Rovnyak. Adv. Texts. Birkhäuser, 1994, xii + 250 pp, $49.50. [ISBN 0-8176-5111-X] Begins with self-contained treatment of harmonic and subharmonic functions; then develops Hardy space theory in the unit disk and in the half-plane. Second half "presents concepts from the theory of univalent functions and Loewner families leading to proofs of the Bieberbach, Robertson, and Milin conjectures." Clear exposition; no exercises. BH

Differential Equations, S*(14). *Differential Equations with Mathematica.* Kevin R. Coombes, *et al.* Wiley, 1995, v + 218 pp, $20.95 (P). [ISBN 0-471-10874-X] Useful supplement to ODE course using *Mathematica.* 89 computing-intensive problems, *Mathematica* instruction and troubleshooting. Written to coordinate with Boyce and DiPrima, but could be used with any text. SK

Differential Equations, P. *Concerning the Hilbert 16th Problem.* Eds: Yu. Ilyashenko, S. Yakovenko. Transl. Ser. 2, V. 165. AMS, 1995, vii + 219 pp, $95. [ISBN 0-8218-0362-X] Qualitative properties of vector fields on the plane. One introductory paper and 5 research papers.

Partial Differential Equations, P. *Inverse Problems in Diffusion Processes.* Eds: Heinz W. Engl, William Rundell. SIAM, 1995, xi + 232 pp, $58 (P). [ISBN 0-89871-351-X] Proceedings of 1994 GAMM–SIAM Symposium.

Partial Differential Equations, P. *Blow-up in Quasilinear Parabolic Equations.* Alexander A. Samarskii. Transl: Michael Grinfeld. Expos. in Math., V. 19. Walter de Gruyter, 1995, xxi + 535 pp, DM 328. [ISBN 3-11-012754-7]

Partial Differential Equations, P. *Third International Conference on Mathematical and Numerical Aspects of Wave Propagation.* Ed: Gary Cohen. SIAM, 1995, xiii + 808 pp, $106 (P). [ISBN 0-89871-350-1] Proceedings of the April 1995 conference.

Partial Differential Equations, P. *Nonlinear Variational Problems and Partial Differential Equations.* Eds: Antonio Marino, M.K. Venkatesha Murthy. Pitman Res. Notes in Math. Ser., V. 320. Longman Scientific & Technical (US Copub: Wiley), 1995, 303 pp, $84.95 (P). [ISBN 0-582-23436-0] Proceedings of a 1990 conference in Italy.

Partial Differential Equations, P. *The Technique of Pseudodifferential Operators.* H.O. Cordes. London Math. Soc. Lect. Note Ser., V. 202. Cambridge Univ Pr, 1995, xii + 382 pp, $34.95 (P). [ISBN 0-521-37864-8]

Partial Differential Equations, P. *Partial Differential Equations in China.* Eds: Chaohao Gu, Xiaxi Ding, Chung-Chun Yang. Math. & Its Applic., V. 288. Kluwer Academic, 1994, x + 181 pp, $80. [ISBN 0-7923-2857-4] Summary of work on PDE's up to early 1980's.

Partial Differential Equations, T(18: 2), S, P. *Positive Harmonic Functions and Diffusion.* Ross G. Pinsky. Stud. in Adv. Math., V. 45. Cambridge Univ Pr, 1995, xvi + 474 pp, $79.95. [ISBN 0-521-47014-5] Draws on both analytic and probabilistic approaches to existence and characterization of positive harmonic functions for second-order elliptic operators. Assumes basic graduate background in analysis, PDE's, and probability. After introducing diffusion processes and spectral theory for elliptic operators, treats Lyapunov method, Martin boundary, Brownian motion, and manifolds. With exercises, notes, references, index. JS

Partial Differential Equations, T(16: 1). *Introductory Applications of Partial Differential Equations, With Emphasis on Wave Propagation and Diffusion.* G.L. Lamb, Jr. Wiley, 1995, xii + 471 pp, $69.95. [ISBN 0-471-31123-5] For a second course in PDE's. Repeats many topics covered in a typical first course, but somewhat more thoroughly. Also covers several special topics, such as Babinet's Principle, oscillations of an inhomogeneous chain, and invariants of linear parabolic equations. JO

Partial Differential Equations, P. *Nonlinear Evolution Equations.* Ed: N.N. Uraltseva. Transl. Ser. 2, V. 164. AMS, 1995, x

+ 220 pp, $95. [ISBN 0-8218-4123-8] 9 papers from the differential equations and mathematical physics seminar at St. Petersburg State University.

Dynamical Systems, P. *Qualitative Theory of Dynamical Systems: The Role of Stability Preserving Mappings.* Anthony N. Michel, Kaining Wang. Pure & Appl. Math., V. 186. Marcel Dekker, 1995, x + 450 pp, $150. [ISBN 0-8247-9420-6] A stability-preserving mapping between dynamical systems preserves trajectories, invariant sets, and their (in)stability. Abstract theory of such mappings, examples from ODE's, PDE's, discrete systems, and more. For engineers and physical scientists. SK

Dynamical Systems, T(16–17). *Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos, Second Edition.* Rüdiger Seydel. Interdisc. Appl. Math., V. 5. Springer-Verlag, 1994, xv + 407 pp, $49. [ISBN 0-387-94316-1] Introduction to bifurcation analysis, with emphasis on computation, for scientists and engineers. Examples and geometrical interpretation replace proof. Revision of the author's 1988 *From Equilibrium to Chaos* (TR, January 1989). SK

Numerical Analysis, P. *Local Uniform Grid Refinement for Time-dependent Partial Differential Equations.* R.A. Trompert. CWI Tract, V. 107. Centrum voor Wiskunde en Informatica, 1995, 160 pp, Dfl. 40 (P). [ISBN 90-6196-452-0]

Operator Theory, P. *Topics in Abstract Differential Equations II.* S. Zaidman. Pitman Res. Notes in Math. Ser., V. 321. Longman Scientific & Technical (Copub: Wiley), 1995, 163 pp, $49.95 (P). [ISBN 0-582-25340-3]

Operator Theory, P. *Multivariable Operator Theory.* Eds: Raúl E. Curto, *et al.* Contemp. Math., V. 185. AMS, 1995, x + 380 pp, $63 (P). [ISBN 0-8218-0298-4] Papers from 1993 conference at University of Washington.

Functional Analysis, T, S, P*, L. *Functional Analysis: Theory and Applications.* R.E. Edwards. Dover, 1995, xiii + 783 pp, $19.95 (P). [ISBN 0-486-68143-2] Preliminary chapter treats set theory, topology. Topics include topological vector spaces, Hahn–Banach theorem, fixed-point theorems, topological duality, applications to partial differential equations. Extensive bibliography. A bargain! BH

Analysis, P. *Opial Inequalities with Applications in Differential and Difference Equations.* Ravi P. Agarwal, Peter Y.H. Pang. Math. & Its Applic., V. 320. Kluwer Academic, 1995, x + 393 pp, $157. [ISBN 0-7923-3365-9]

Analysis, P. *Approximation Theory, Wavelets and Applications.* Eds: S.P. Singh, Antonio Carbone, B. Watson. NATO ASI Ser. C, V. 454. Kluwer Academic, 1995, xxiii + 572 pp, $254. [ISBN 0-7923-3334-9] Proceedings of a 1994 NATO Advanced Study Institute in Italy.

Analysis, T*(17: 2), P. *Integration and Probability.* Paul Malliavin. Grad. Texts in Math., V. 157. Springer-Verlag, 1995, xxi + 322 pp, $44. [ISBN 0-387-94409-5] An interesting alternative to the usual first-year graduate analysis text. After a compact but thorough treatment of standard topics, introduces such unusual applications as spectral analysis, Sobolev spaces, pseudo-differential operators, probability theory, martingales, and stochastic calculus of variations. Translated from French. DB

Differential Geometry, T(18), L. *Geometric Measure Theory: A Beginner's Guide, Second Edition.* Frank Morgan. Academic Pr, 1995, ix + 175 pp, $34.95. [ISBN 0-12-506857-3] Intended to "make the standard text, *Geometric Measure Theory*, by H. Federer, more accessible," a laudable goal. Includes a chapter on soap bubble clusters. (*First Edition*, TR, January 1989; Extended Review, October 1989.) JO

Differential Geometry, T(17–18: 1, 2). *Differential and Riemannian Manifolds.* Serge Lang. Grad. Texts in Math., V. 160. Springer-Verlag, 1995, xiii + 364 pp, $49.50. [ISBN 0-387-94338-2] Arguably appropriate as a text for an introductory graduate course, but like many of the author's books, better used as a reference. Introduces differential topology and differential geometry, and includes results from Riemannian geometry and calculus of variations. JO

Geometry, T(18: 1, 2), L. *Riemannian Geometry, Second Revised Edition.* Wilhelm P.A. Klingenberg. Stud. in Math., V. 1. Walter de Gruyter, 1995, x + 409 pp, DM 158. [ISBN 3-11-014593-6] Excellent introduction to Riemannian manifolds. Includes a chapter on differential geometric foundations, but this is not an introductory text in differential geometry. (*First Edition*, TR, November 1984; Extended Review, August–September 1985.) JO

Geometry, P, L**.** *Kaleidoscopes: Selected Writings of H.S.M. Coxeter.* F. Arthur Sherk. Canadian Math. Soc. Ser. of Mono. & Adv. Texts. Wiley, 1995, xxx + 439 pp, $74.95. [ISBN 0-471-01003-0] 26 papers selected by Coxeter from his many papers on reflection groups. A tribute both to the Canadian Mathematical Society on its 50th anniversary, and to one of its founding members. JNC

Optimization, P. *Variational Inequalities and*

Network Equilibrium Problems. Eds: F. Giannessi, A. Maugeri. Plenum Pr, 1995, xii + 305 pp, $89.50. [ISBN 0-306-45007-0] Proceedings of a 1994 Summer School in Italy.

Optimal Control, P. *Problems and Methods of Optimal Control.* Leonid D. Akulenko. Math. & Its Applic., V. 286. Kluwer Academic, 1994, xv + 342 pp, $150.50. [ISBN 0-7923-2855-8]

Stochastic Processes, T(16–17: 1, 2), L. *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling.* Randolph Nelson. Springer-Verlag, 1995, xxviii + 583 pp, $49. [ISBN 0-387-94452-4] Basic introduction to probability theory and stochastic processes. Examples and applications drawn from queueing theory and computer systems. Assumes calculus, linear algebra. Topics include moment-generating functions, Laplace transforms, renewal processes, queueing networks, matrix geometric solutions. LB

Stochastic Processes, T(18: 1, 2), S, P. *Diffusions, Markov Processes, and Martingales, Volume 1: Foundations, Second Edition.* L.C.G. Rogers, David Williams. Wiley, 1994, xx + 386 pp, $79.95. [ISBN 0-471-95061-0] 3 long chapters treat motivation, background, and more detailed topics. Chapter 1, on Brownian motion, is updated. Chapter 2, substantially expanded, now reviews necessary background from measure theory, probability, stochastic processes. Chapter 3 is little changed from *First Edition.* With companion volume on Itô calculus, prepares reader for research in the area. Infectiously enthusiastic style. LB

Stochastic Processes, P. *Topics in the Constructive Theory of Countable Markov Chains.* G. Fayolle, V.A. Malyshev, M.V. Menshikov. Cambridge Univ Pr, 1995, 169 pp, $44.95. [ISBN 0-521-46197-9] Provides methods, based on Lyapunov functions, giving necessary and sufficient conditions for determining when a Markov chain is ergodic, null recurrent, or transient. Deals primarily with deflected random walks in $\mathbb{Z}_+^N$. RSK

Stochastic Processes, P. *Stochastic Geometry Models in Image Analysis and Spatial Statistics.* M.N.M. van Lieshout. CWI Tract, V. 108. Centrum voor Wiskunde en Informatica, 1995, 172 pp, Dfl. 40 (P). [ISBN 90-6196-453-9]

Elementary Statistics, T(13–15: 1), L. *Geographical Data Analysis.* Nigel Walford. Wiley, 1995, xi + 446 pp, $59.95 (P). [ISBN 0-471-94162-X] Elementary statistics in the context of geography. Case study format. 4 sample research projects (on land-use, weathering of chalk outcroppings, river pollution, village communities) from formulation of hypotheses through data collection and analysis. Projects demonstrate use of MINITAB and SPSS–X. LB

Elementary Statistics, T*(13: 1). *The Basic Practice of Statistics.* David S. Moore. WH Freeman, 1995, xviii + 680 pp. [ISBN 0-7167-2628-9] Text for a data-oriented statistics course; author is a leader of the reform movement in teaching statistics. Emphasizes statistical thinking, with less theory and fewer recipes. User-friendlier than Moore and McCabe's *Introduction to the Practice of Statistics* (*First Edition*, TR, June–July 1989). RSK

Elementary Statistics, T*(13–14: 1, 2). *Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery.* Ronald N. Forthofer, Eun Sul Lee. Academic Pr, 1995, xviii + 567 pp, $59.95. [ISBN 0-12-262270-7] Emphasizes context and applicability as well as methods. Uses real data; calculations are illustrated with MINITAB (SAS and Stata commands appear in an appendix). Includes a chapter on analysis of follow-up life tables, material on multiple and logistic regression, and on tolerance and prediction intervals. RSK

Statistical Methods, S(16–18), P, L. *Practical Methods for Design and Analysis of Complex Surveys.* Risto Lehtonen, Erkki J. Pahkinen. Wiley, 1995, ix + 337 pp, $55.95. [ISBN 0-471-93934-X] Sampling techniques, use of auxiliary information, handling missing data, variance and ratio estimation, one- and two-way table analysis, and multivariate survey analysis. Includes case studies. RWJ

Statistics, T(15–17: 2). *Introduction to Statistical Inference.* E.S. Keeping. Dover, 1995, xi + 451 pp, $15.95 (P). [ISBN 0-486-68502-0] Topics include probability, summarizing data, testing hypotheses, sampling procedures, nonparametric methods, bivariate problems, and regression. Assumes some differential/integral calculus. Republication of 1962 Van Nostrand edition. RWJ

Mathematical Computing, S(13–14). *Practical Mathematics Using MATLAB.* Gunnar Backstrom. Studentlitteratur, 1995, 165 pp, (P). [ISBN 91-44-49231-6] Collection of examples and exercises that can be studied using MATLAB. Includes topics from calculus (e.g., computing difference quotients, plotting derivatives), statistics (e.g., data fitting), and differential equations (e.g., numerical methods). Teaches simple MATLAB programming. LC

Computer Systems, P. *When You Can't*

Find Your UNIX System Administrator. Linda Mui. O'Reilly & Assoc, 1995, xiii + 139 pp, $17.95 (P). [ISBN 1-56592-104-6]

Computer Systems, P. *The USENET Handbook: A User's Guide to Netnews.* Mark Harrison. O'Reilly & Assoc, 1995, xiv + 372 pp, $24.95 (P). [ISBN 1-56592-101-1]

Computer Systems, S(15–16), C, P. *Graphics Gems V.* Ed: Alan W. Paeth. Academic Pr, 1995, xx + 438 pp, $49.95 (P), with disk. [ISBN 0-12-543457-X]

Computer Systems, P. *Using Email Effectively.* Linda Lamb, Jerry Peek. O'Reilly & Assoc, 1995, xii + 146 pp, $14.95 (P). [ISBN 1-56592-103-8]

Computer Systems, P. *HTML Pocket Reference.* Specialized Systems Consultants, 1995, 14 pp, $4.50 (P). [ISBN 0-916151-79-4]

Computer Science, C, P. *Advanced Algorithms for Neural Networks: A C++ Sourcebook.* Timothy Masters. Wiley, 1995, xiv + 431 pp, $44.95 (P), with disk. [ISBN 0-471-10588-0]

Computer Science, P. *Domain-Based Parallelism and Problem Decomposition Methods in Computational Science and Engineering.* Eds: David E. Keyes, Youcef Saad, Donald G. Truhlar. SIAM, 1995, xvi + 323 pp, $45 (P). [ISBN 0-89871-348-X] 17 articles emphasizing transferable algorithmic techniques.

Computer Science, P. *High Performance Computing: Problem Solving with Parallel and Vector Architectures.* Ed: Gary W. Sabot. Addison-Wesley, 1995, xvi + 246 pp. [ISBN 0-201-59830-2] Case studies from several problem domains illustrate programming techniques for high-performance computers.

Computer Science, P. *Quality of Communication-Based Systems.* Ed: Günter Hommel. Kluwer Academic, 1995, x + 195 pp, $129. [ISBN 0-7923-3259-8] Proceedings of a 1994 workshop in Berlin. Organized into sections on specification techniques, quantitative modeling, and distributed systems organization.

Applications (Fluid Mechanics), P. *Scaling Phenomena in Fluid Mechanics.* Grigory Isaakovich Barenblatt. Cambridge Univ Pr, 50 pp, $5.95 (P). [ISBN 0-521-46920-1]

Applications (Fluid Mechanics), P. *Nontraditional Methods in Mathematical Hydrodynamics.* O.V. Troshkin. Transl. of Math. Mono., V. 144. AMS, 1995, viii + 197 pp, $98. [ISBN 0-8218-0285-2]

Applications (Mechanics), P. *Asymptotic Theories for Plates and Shells.* R.P. Gilbert, K. Hackl. Pitman Res. Notes in Math. Ser., V. 319. Longman Scientific & Technical (US Copub: Wiley), 1995, 131 pp, $50 (P). [ISBN 0-582-24875-2] 9 papers, most based on presentations at the SIAM 40th Anniversary Meeting (Los Angeles, 1992).

Applications (Physics), S(16), P. *A Student's Guide to Fourier Transforms with Applications in Physics and Engineering.* J.F. James. Cambridge Univ Pr, 1995, x + 131 pp, $15.95 (P). [ISBN 0-521-46829-9] Concise introduction "using qualitative arguments wherever possible, and avoiding unnecessary mathematics." Focuses sharply on applications, including optics, spectroscopy, electronics, telecommunications, etc. Many nice examples and figures. BH

Applications (Quantum Theory), P, L. *Quantum Chaos Between Order and Disorder.* Giulio Casati, Boris Chirikov. Cambridge Univ Pr, 1995, xv + 683 pp, $100. [ISBN 0-521-43291-X] A collection of papers, some previously published, on quantum chaos.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–204: Logical Approach to Systems Theory.* Shingo Takahashi, Yasuhiko Takahara. Springer-Verlag, 1995, xi + 174 pp, $43 (P). [ISBN 3-540-19956-X]

Applications (Systems Theory), T(17–18: 1, 2), S, P. *An Introduction to Infinite-Dimensional Linear Systems Theory.* Ruth F. Curtain, Hans Zwart. Texts in Appl. Math., V. 21. Springer-Verlag, 1995, xviii + 698 pp, $49. [ISBN 0-387-94475-3] Treats spectral systems on Hilbert space to minimize prerequisites. Topics (Cauchy problem, controllability, observability, stabilizability, detectability, Hankel operators, relationship between state-space and frequency domain) are illustrated with three main examples: diffusion equation, wave equation, and a delay equation. RM

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–203: Macrosystems Theory and its Applications: Equilibrium Models.* Yury S. Popkov. Springer-Verlag, 1995, xiii + 323 pp, $69 (P). [ISBN 3-540-19955-1]

Reviewers

LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; LC: Laura Chihara, St. Olaf; BH: Bruce Hanson, St. Olaf; RWJ: Roger W. Johnson, Carleton; SK: Steve Kennedy, Carleton; RSK: Richard S. Kleber, St. Olaf; LCL: Loren C. Larson, St. Olaf; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; JS: John Schue, Macalester; KES: Kay E. Smith, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; MW: Martha Wallace, St. Olaf; PZ: Paul Zorn, St. Olaf.

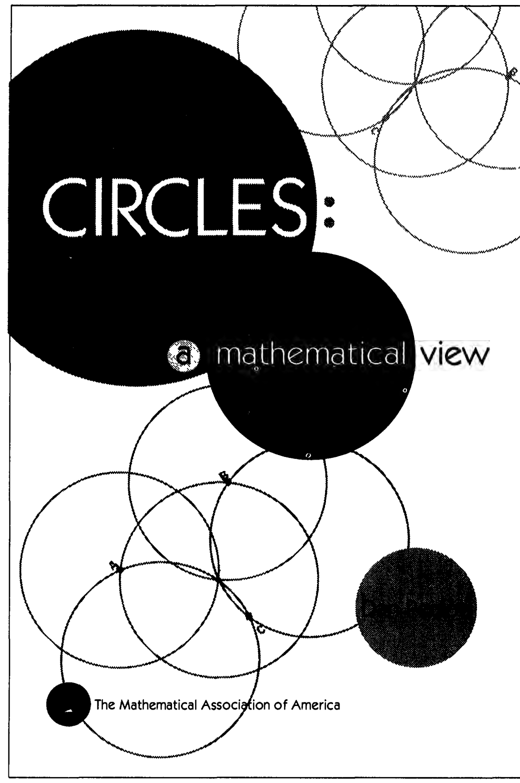
CIRCLES
A Mathematical View

Dan Pedoe

Revised edition

Illuminates the fundamental aspects of college geometry, non-Euclidean geometry, and other branches of mathematics where the circle plays an important role.

No school or undergraduate mathematics library should be without this book.
London Times Education Supplement



This revised edition of a mathematical classic originally published in 1957 will bring to a new generation of students the enjoyment of investigating that simplest of mathematical figures, the circle. The author has supplemented this new edition with a special chapter designed to introduce readers to the vocabulary of circle concepts with which the author could assume his readers of two generations ago were familiar. For example, Pedoe carefully explains what is meant by the *circumcircle, incircle,* and *excircles* of a triangle as well as the *circumcentre, incentre,* and *orthocentre.* The reader is then well equipped to understand his discussions of the nine-point circle and of Feuerbach's theorem. In an appendix, Pedoe includes a biographical sketch of Karl Wilhelm Feuerbach, a little known mathematician with a tragically short life, who published his theorem in a slender geometric treatise in 1822.

Readers of *Circles* need only be armed with paper, pencil, compass, and straightedge to find great pleasure in following the constructions and theorems. Those who think that geometry using Euclidean tools died out with the ancient Greeks will be pleasantly surprised to learn many interesting results which were only discovered in modern times. And those who think that they learned all they needed to know about circles in high school will find much to enlighten them in chapters dealing with the representation of a circle by a point in three-space, a model for non-Euclidean geometry, and the isoperimetric property of the circle.

144 pp., Paperbound, 1995 ISBN-0-88385-518-6
List: $18.95 MAA Member: $14.50
Catalog Code: CIRCLES/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

Qty.	Catalog Code	Price	Amount
_____	CIRCLES/JR	_____	_____
_____	__________	_____	_____
		TOTAL	_____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

THE AMERICAN MATHEMATICAL MONTHLY

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

Volume 103, Number 2 February 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

Richard BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

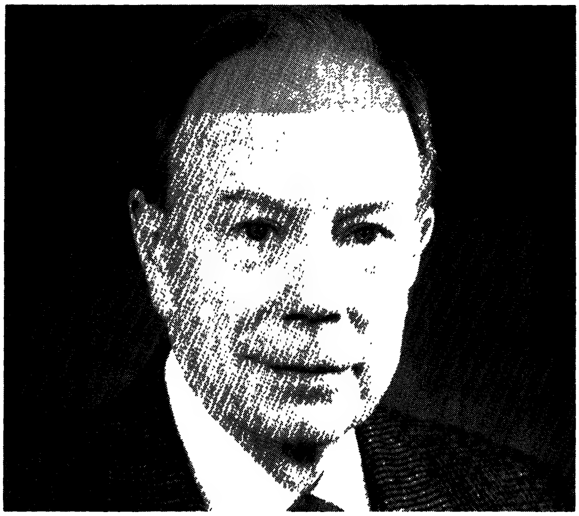
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Yueh-Gin Gung and Dr. Charles Y. Hu Award for Distinguished Service to Andrew Gleason

H. O. Pollak



Andrew M. Gleason was born in Fresno, California in 1921, received most of his schooling in Yonkers, New York, and went to Yale, where he was in the top five in the Putnam Competition for three years in a row. After four years of active and fortunately mathematical duty in the Navy, he became a Junior Fellow at Harvard in 1946. With some time out for further active duty during the Korean War, he has been at Harvard ever since, becoming a full professor in 1957. His wife, Jean Berko Gleason, is Professor of Psychology at Boston University; they have three daughters, Katherine, Pamela, and Cynthia.

In thinking about, and admiring, Andy Gleason's career, your natural reference is the total profession of a mathematician: designing and teaching courses, advising on education at all levels, doing research, consulting for the users of mathematics, acting as a leader of the profession, cultivating mathematical talent, and serving one's institution. Andy Gleason is that rare individual who has done all of these superbly.

His influence on mathematics education has covered over 40 years, and has been outstanding. He has been heavily involved in the thinking about mathematics education ever since the 1950's. A few examples: He was chairman of the Advisory Board to the School Study Mathematics Group; he organized the Cambridge Conference on School Mathematics; and he was on the Advisory Board to USMES, the Unified Science and Mathematics program in the Elementary Schools. All of these, while controversial, were influential and thoughtful projects. He

taught some elementary school math himself, and pioneered reporting what didn't work as well as what did. He was the chief mathematical advisor to Houghton Mifflin and the Dolciani Series for many years, and undertook many key tasks nationally, participating, for example, in the thinking that led to the establishment of the Mathematical Sciences Education Board on which he served during its first four years of existence. In recent years, Andy's "guiding hand," as Anneli and Peter Lax have described it, can be seen throughout Harvard Project Calculus. He has continued to be active in education since his 1992 retirement, working with the Massachusetts State Board of Education on curricular reform, and with the Interactive Mathematics Project as a member of its Advisory Board.

The students at Harvard University, where Gleason was the Hollis Professor of Mathematics and Natural Philosophy, have been the beneficiaries of many other curricular innovations besides Project Calculus. For example, he designed a second-year calculus experience truly integrated with linear algebra (the students took naturally to a rather abstract development), and developed Natural Sciences 1a (Euclid, Archimedes, and Newton, inter alia) as part of the General Education Program at Harvard. Both of these show his strong sense of history in the teaching of mathematics and science.

In 1962 Gleason was the MAA's Earle Raymond Hedrick Lecturer at the summer meetings in Vancouver, B.C. In addition, he has served the Association as a member of a number of committees, including the Committee on the Putnam Prize Competition, the Committee on the Hedrick Lectures, the Science Policy Committee, and the Development Committee.

His success in mathematical research has been outstanding. He was one of the major contributors to finishing the solution of Hilbert's Fifth Problem, and his research has had major influence in areas as apparently disparate as quantum mechanics and combinatorics. He has been known to explain that his strength is a thorough knowledge of the fundamentals, and he rarely, if ever, turns down a good problem. Gleason is a member of the National Academy of Sciences, a past president of the American Mathematical Society, and was both chairman of the organizing committee and president for the International Congress of Mathematicians at Berkeley in 1986. His mathematical tastes are very broad. Although he is classified as an abstract analyst, and this field encompasses the majority of his research papers, he has worked, and supervised dissertations, in many fields of both pure and applied mathematics. The one I remember best is the thesis in algebraic coding theory of Jessie MacWilliams, whom he helped to develop into one of the outstanding women mathematicians of our time. Many of his students agree that his quickness in thought and understanding make him a tough supervisor. He once characterized his main function in working with future Ph.D.'s as giving them the opportunity to find out how good they are!

Gleason participated in a group on coding theory, which met monthly for about ten years, and which included Ed Assmus, H. F. Mattson, Jr., John N. Pierce, Vera Pless and Gene Prange among its "regulars." Gleason's interest in coding theory and cryptography extends beyond his own research, and that of a number of his students, to his consulting both for industry and for the nation's intelligence and security programs for over 50 years. He worked with both NSA and IDA, and consulted at many levels. In summary, as David Lieberman has put it: "It has been inspiring to review and comprehend the role which he has played in shaping the science, the teaching, and the application of mathematics, both through his own contributions and through the many lives and careers he has so strongly influenced."

Using Self-Similarity to Find Length, Area, and Dimension

James T. Sandefur

As a high-school student many years ago, I saw my first application of self-similarity. Specifically, my teacher found the fractional equivalent of $x = 0.33\ldots$ by noticing that

$$x = 0.333\ldots = 0.3 + 0.0333\ldots = 0.3 + (0.1)(0.333\ldots) = 0.3 + 0.1x.$$

Solving gave that $x = 1/3$. In college, my mathematics professor observed that $x = 0.999\ldots = 0.9 + 0.1x$. Solving for x showed that

$$0.999\ldots = 1.$$

In teaching calculus, many of us have summed convergent infinite geometric series by letting $x = 1 + r + r^2 + \ldots = 1 + rx$ and solving for x to get that

$$1 + r + r^2 + \ldots = 1/(1 - r).$$

The stratagem used above was to notice that the number x involves a fractional part of itself, and then use this self-similarity to develop an equation involving the number x. This stratagem gives the sum of a convergent series, but does not prove that the series converges.

Number theorists often use this technique to compute the limit of continued fractions. For example, let

$$x = \cfrac{1}{2 + \cfrac{1}{2 + \ldots}} = \frac{1}{2 + x}.$$

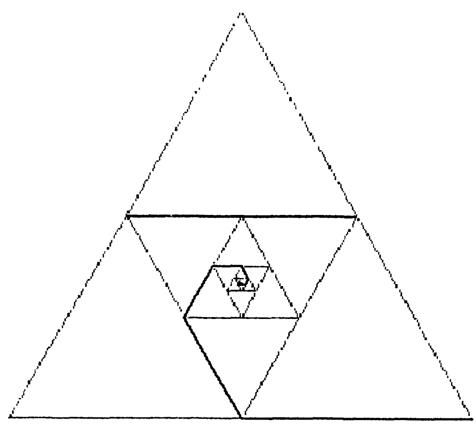
Solving gives that $x = \sqrt{2} - 1$.

To motivate my students when teaching geometric series, I try to study models of situations in which a geometric series arises naturally. One example that I started using was to generate spirals and then find their lengths and areas. One advantage to these problems is that the algebraic observation described above becomes graphically obvious. Another advantage is the derivation of series involving roots and trigonometric functions. Another example in which series arise is the computation of the area of fractals. This article will study the interplay between these examples of series and the little algebraic trick seen above, as well as show how this algebraic use of self-similarity can motivate the concept of fractal dimension.

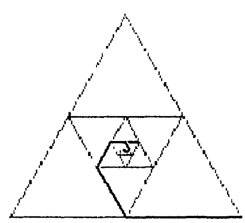
A term needed in the following sections is **scaling factor**. Specifically, suppose a side of one figure is of length L and the corresponding side of a second similar figure is of length l. Then the **scaling factor** of the second figure to the first is defined to be

$$p = l/L.$$

LENGTHS OF SPIRALS. In Figure 1(a) is an equilateral triangle with sides of length one. The midpoints of each side are connected to form an inscribed equilateral triangle. The midpoints of each side of the inscribed triangle are connected giving another inscribed equilateral triangle. The process is continued, giving a series of nested equilateral triangles. Half of one side of each triangle is highlighted, forming a spiral. What is the length of this spiral?



(a)



(b)

Figure 1. Figure 1(a) is a spiral inside an equilateral triangle with sides of length 1. Figure 1(b) is formed by deleting the outer triangle from Figure 1(a).

Erasing the original outer triangle and rotating the remaining figure gives Figure 1(b). Note that this second figure is similar to the original, except the sides are of length 0.5. Thus the scaling factor of the second figure to the first is $p = 1/2$. Let S and s represent the lengths of the spirals in Figures 1(a) and (b), respectively. Since Figures 1(a) and (b) are similar, the length of the spiral in Figure 1(b) is **half** the length of the spiral in Figure 1(a), that is, $s = 0.5S$. But the spiral in Figure 1(b) resulted by deleting the first branch of the spiral in Figure 1(a), that is, $S = s + 0.5$. Substitution gives that $S = 0.5S + 0.5$, so $S = 1$. Thus, we have found that the spiral in Figure 1(a) is the same length as one side of the original equilateral triangle.

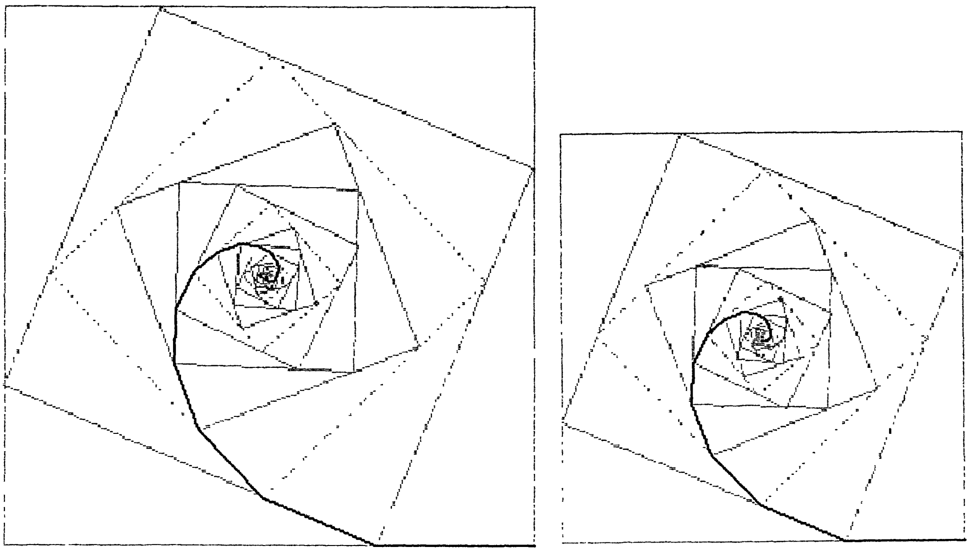
Let's connect this geometric approach to finding the length of a spiral to the algebraic one in the Introduction. Note that the length of the spiral is also given by the infinite geometric series $S = 0.5 + 0.5^2 + 0.5^3 + \dots$, and we obtain the same equation, that is,

$$S = 0.5 + 0.5^2 + 0.5^3 + \dots = 0.5 + 0.5(0.5 + 0.5^2 + \dots) = 0.5 + 0.5S.$$

For the next example, divide each side of a unit square into two parts of lengths a and b, where $a + b = 1$. Connect the division points, forming a new square with sides of length

$$h = \sqrt{a^2 + b^2}.$$

Divide each side of this new square into 2 parts of length ah and bh and connect the division points. Continue dividing the sides of the new squares into the proportions of a to b, giving Figure 2(a). The a portion of one side of each square is highlighted, giving the spiral in that figure. As before, delete the outer square, giving a similar figure, Figure 2(b), with outer sides of length $h = \sqrt{a^2 + b^2}$, which is the scaling factor between the two figures.



(a) (b)

Figure 2. Figure 2(a) is a spiral formed inside a square with sides of length one. Figure 2(b) is formed by deleting the outer square from Figure 2(a).

If we designate the length of the spirals as S and s, as before, then the relationship implied by being similar figures is $s = hS = \sqrt{a^2 + b^2}\,S$. But since the larger spiral is made up of the smaller spiral and part of a side of the outer square, we have $S = s + a$. Substitution gives $S = \sqrt{a^2 + b^2}\,S + a$ and solving gives

$$S = \frac{a}{1 - \sqrt{a^2 + b^2}} = \frac{a\left(1 + \sqrt{a^2 + b^2}\right)}{1 - a^2 - b^2}.$$

Using the facts that $1 - a^2 - b^2 = 2ab$ and $a = 1 - b$ gives

$$S = \frac{a\left(1 + \sqrt{a^2 + b^2}\right)}{2ab} = \frac{1 + \sqrt{2b^2 - 2b + 1}}{2b}.$$

This answer could also be obtained by generating the geometric series for the length of the spiral.

Letting $b = 1/2$ gives a spiral of length $1 + \sqrt{0.5}$ which is also the perimeter of the outer corner triangle. An interesting exercise is to explain geometrically why these lengths are the same.

Solving for b in terms of S gives

$$b = \frac{2S - 1}{2S^2 - 1}$$

with $a = 2S(S - 1)/(2S^2 - 1)$. If you want to generate a spiral of length $S = 2$, substitution gives that $b = 3/7$ and $a = 4/7$. In fact, if S is an integer, than the numerators of a and b, together with $2S^2 - 2S + 1$ form Pythagorean triples.

Letting b go to one gives a smooth spiral of length 1, the length of a side of the original square. The interested reader might try to develop a differential equation for the limit curve and find the solution curve.

An interesting geometric series arises in finding the length of spirals constructed in regular n-gons. The idea is to connect the midpoints of each of the sides of the n-gon, generating an inscribed n-gon. Repeating this procedure generates a sequence of inscribed n-gons. Connect half of one side of each n-gon to generate a spiral. See the pentagon in Figure 3 for one example.

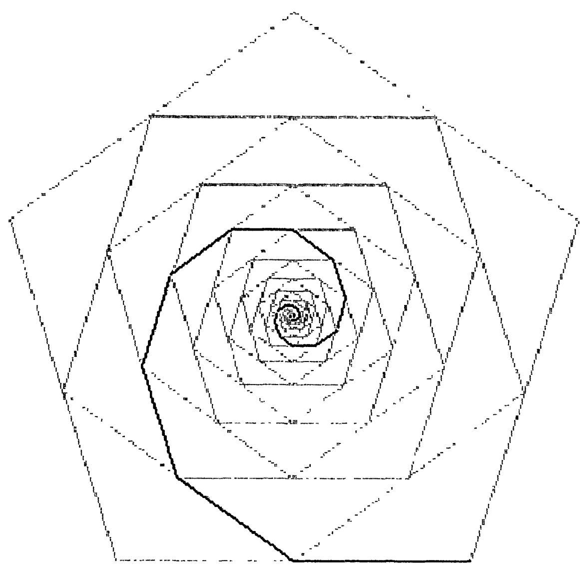


Figure 3. A spiral formed inside a regular pentagon.

Recall that the exterior angle of an n-gon, that is, the angle between the extension of one side of the n-gon and an adjacent side of the n-gon, is $\theta = 2\pi/n$. Connect the midpoints of the sides of the original n-gon to obtain a similar n-gon. Suppose the sides of the original n-gon are of length L. Each of the new sides form the base of an isosceles triangle with two sides of length $L/2$, and with base angles that are π/n, half the exterior angle. See Figure 4 where $n = 5$. Simple

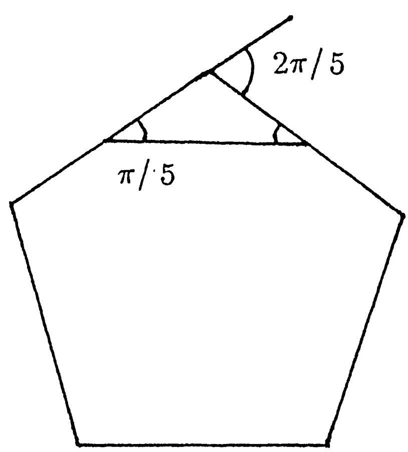


Figure 4. A pentagon with the exterior angle and the base angles of an isosceles triangle marked.

trigonometry gives that the length of the sides of the inscribed n-gon is

$$l = L\cos(\pi/n).$$

Because the length of a side of the smaller n-gon is $\cos(\pi/n)$ times the length of a side of the larger n-gon, then the scaling factor of the inscribed n-gon to the outer n-gon is $p = \cos(\pi/n)$.

Let S represent the length of the spiral resulting from constructing inscribed n-gons and connecting one side of each. See Figure 3 for example. Erasing the outer n-gon, say the outer pentagon in Figure 3, gives a similar figure with scaling factor $\cos(\pi/n)$. Denote the length of the spiral in this figure as s. Thus,

$$s = \cos(\pi/n)S.$$

But the second spiral is just the first spiral with the first edge removed, that is,

$$S = L/2 + s.$$

Thus, the length of the original spiral is

$$S = \frac{L}{2[1 - \cos(\pi/n)]}.$$

Note that we have only computed the infinite geometric series

$$(L/2)\sum_{i=0}^{\infty}[\cos(\pi/n)]^i.$$

An interesting problem is to find the length of a spiral determined by consecutively inscribed n-gons which are constructed by dividing each side of the previous n-gon in the proportions of a to b and connecting consecutive points, as was done for the square.

AREA OF SPIRALS. To compute the area of a spiral, you only need to remember that if two figures are similar with scaling factor p, then their areas are related by p^2.

When considering the inscribed regular n-gons, shade one corner of each inscribed figure instead of just one line. This has been done for the pentagon in Figure 5. Let A represent the area of this spiral. Erase the outer n-gon giving a new, similar figure with scaling factor $\cos(\pi/n)$. Call the area of this spiral a. Then

$$a = A\cos^2(\pi/n).$$

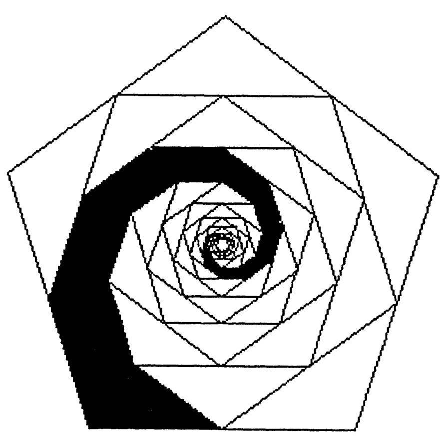


Figure 5. A spiral inside a regular pentagon.

Simple computations give that the area of the deleted triangle is

$$T = (L^2/4)\sin(\pi/n)\cos(\pi/n).$$

Since $A = T + a$, we can solve and find that the area of the spiral is

$$A = (L^2/4)\cot(\pi/n).$$

This problem could again be done by summing the geometric series

$$T\sum_{i=0}^{\infty}\left[\cos^2(\pi/n)\right]^i.$$

A simple observation indicates that the n-gon is made up of n of these spirals. Thus a bonus from our calculations is that the area of a regular n-gon with sides of length L is

$$n(L^2/4)\cot(\pi/n).$$

A simpler approach to computing the area of regular n-gons is to construct only one inscribed n-gon and then use the same method, relating the two areas using similarity, but noticing that the difference between the two areas is the area of the n triangles in the corners. You can also compute the length of the diagonals of regular n-gons by using the similarity between the original regular n-gon and the inscribed n-gon. You must consider two separate cases: when n is even and when n is odd.

AREA OF THE KOCH SNOWFLAKE. Many readers are familiar with the Koch snowflake. To construct one side of the Koch snowflake, draw a line of length L. Then construct an equilateral triangle on the middle third of the line and erase the base of the triangle. Repeat this process on each of the four line segments. Keep repeating this process on the smaller line segments; the limit is the Koch curve. Do this for each side of an equilateral triangle with sides of length L; the figure bounded by the three Koch curves is the Koch snowflake, seen in Figure 6.

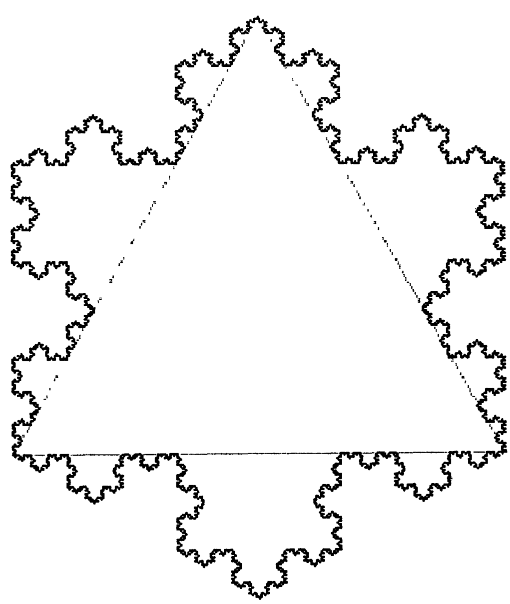


Figure 6. A Koch snowflake formed about an equilateral triangle with sides of length L.

One method for finding the area of the Koch snowflake is to construct the geometric series that sums the areas of all the equilateral triangles that are constructed at each stage of the generation of the Koch curves.

Let us find the area of the Koch snowflake by using the self-similarity of the Koch curve. There are two areas to be determined. The first is the area of the equilateral triangle with sides of length L. This is

$$A = \sqrt{3}L^2/4.$$

The second is the area between the triangle and the Koch curve. Figure 7(a) is a magnification of the region bounded by one side of the triangle and the Koch curve. Call its area x. Then the area of the Koch snowflake is

$$A + 3x.$$

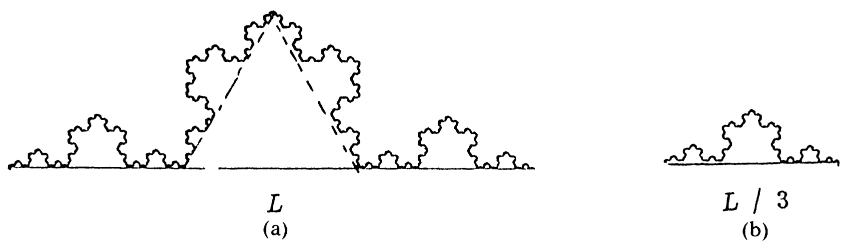


Figure 7. Figure 7(a) is the Koch curve formed on one side of the equilateral triangle in Figure 6. Figure 7(b) is one part of Figure 7(a).

There are two parts to finding the area x. The first part is to compute the area of the equilateral triangle in the middle. Since it is similar to the original equilateral triangle, but with sides whose lengths are one-third the original, the scaling factor is $1/3$ and the area of the triangle is therefore

$$(1/3)^2 A = A/9.$$

The second part is to find the area of the four congruent parts around the triangle, one of which is given in Figure 7(b). Denote the area of the region in Figure 7(b) by y. Then we have that

$$x = A/9 + 4y.$$

Now use self-similarity. The region given in Figure 7(b) is similar to the region in Figure 7(a), but with a scaling factor of one-third. Thus the area of the region in Figure 7(b) is one-ninth the area of the region in Figure 7(a); that is,

$$y = x/9.$$

Substitution gives that

$$x = A/9 + 4y = A/9 + 4x/9.$$

Solving for x gives

$$x = A/5.$$

Thus, the area of the Koch snowflake is

$$A + 3x = A + 3A/5 = 8A/5 = 2\sqrt{3}L^2/5.$$

I have found this approach to be much simpler and more satisfying than the algebraically messy method of computing the geometric series for the area of the Koch snowflake. As word of caution, when generating a fractal, such as the Koch snowflake, the first, second, third and so forth, iterates, are not self-similar. Thus, this technique cannot be used to find the length or area of the approximates to the

fractal. For more details about the Koch snowflake, see Sandefur [3], Peitgen et al. [2], and Peak and Frame [1].

VARIATIONS ON THE KOCH SNOWFLAKE. A variation on the construction of the Koch snowflake is to construct an equilateral triangle with sides of length L, then divide each side in the proportions $p:q:p$, where $p+q+p=1$ and $2p>q$. Construct an isosceles triangle on the middle line segment of each side with the two equal sides being the same length as the two outer portions of the line; that is, of length pL. Erase the base of the isosceles triangle. Repeat this process on each of the four line segments. Keep repeating this process and the limit figure is a variation on the Koch snowflake. Figure 8 is one such curve where $p=0.26$. Figure 9 is another variation with $p=0.4$.

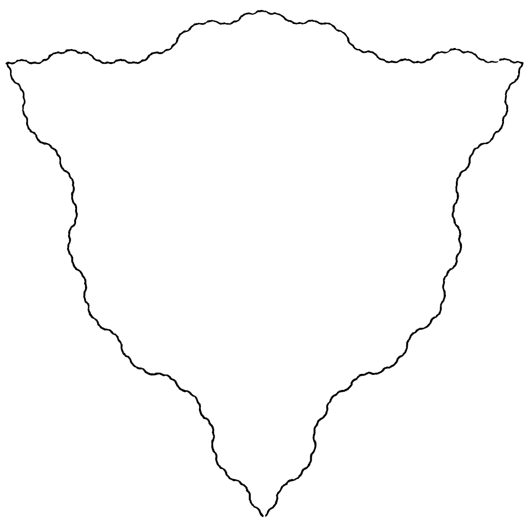


Figure 8. A Koch snowflake in which each side was divided in the proportions 0.26 : 0.48 : 0.26.

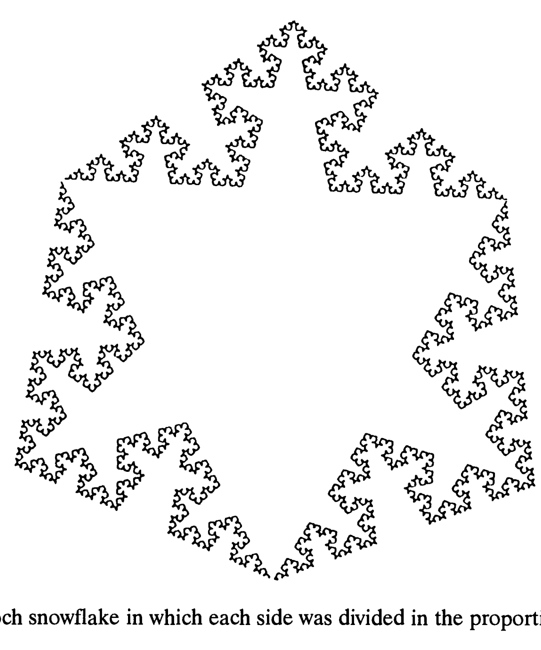


Figure 9. A Koch snowflake in which each side was divided in the proportions 0.4 : 0.2 : 0.4.

There are two parts to computing the area of this Koch snowflake variation. The first is to observe that the area of the equilateral triangle is $A = \sqrt{3}L^2/4$ as before. The second is to find the area bounded by the Koch-like curves and the triangle. Letting x represent the area bounded by one side of the equilateral triangle and the Koch-like curve, we again have that the area of the snowflake is

$$A + 3x.$$

To find x, there are again two parts. The first part is to compute the area of the isosceles triangle, which reasonably simple calculations give as

$$T = \frac{qL^2\sqrt{1-2q}}{4}.$$

The second part is to find the area of each of the four congruent parts around the isosceles triangle. Denote the area of one of these regions by y. Then as before $x = T + 4y$. We note that each of the four smaller regions is similar to the region bounded by the Koch-like curve and one side of the equilateral triangle, but with a scaling factor of p. Thus,

$$y = p^2 x.$$

Substitution gives $x = T + 4y = T + 4p^2x$ or that

$$x = T/(1 - 4p^2).$$

But $1 - 4p^2 = (1 - 2p)(1 + 2p) = q(2 - q)$ so

$$x = \frac{T}{q(2-q)} = \frac{qL^2\sqrt{1-2q}}{4q(2-q)} = \frac{L^2\sqrt{1-2q}}{4(2-q)}.$$

The area of the Koch snowflake variation is

$$A + 3x = \frac{\sqrt{3}L^2}{4} + \frac{3L^2\sqrt{1-2q}}{4(2-q)}.$$

In Figure 10 is another variation of the snowflake. This is the same construction as in Figure 8, but cut out of the triangle instead of built on. This construction can

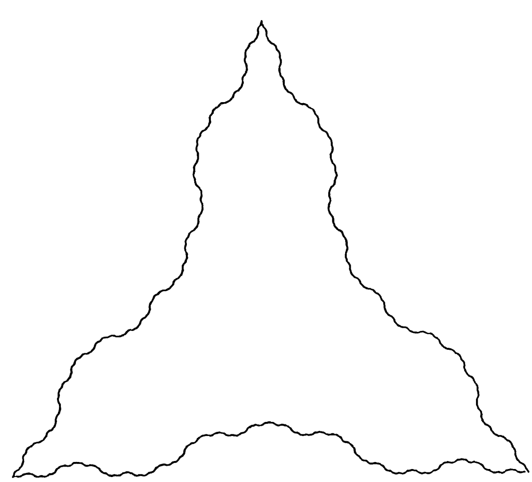


Figure 10. A Koch snowflake where the sides were divided in the proportions of 0.26 : 0.48 : 0.26 and the Koch-like curve was constructed inside the equilateral triangle.

be done for $1/4 < p < 1/3$. The area of the snowflake is then $A - 3x$ instead of $A + 3x$.

THE AREA OF A SKYLINE. Construct a line of length 1. Divide it into the proportions of $p_1 : p_2 : p_3$ (going from left to right) where $p_1 + p_2 + p_3 = 1$. Construct a square on the middle portion. Now divide each of the three **horizontal** line segments into the same three proportions and construct squares on the middle portions. Figure 11(a) is the fractal that is the limit of this process when $p_1 = 0.3$, $p_2 = 0.5$, and $p_3 = 0.2$. One problem is to find the area A of this "skyline"-type figure.

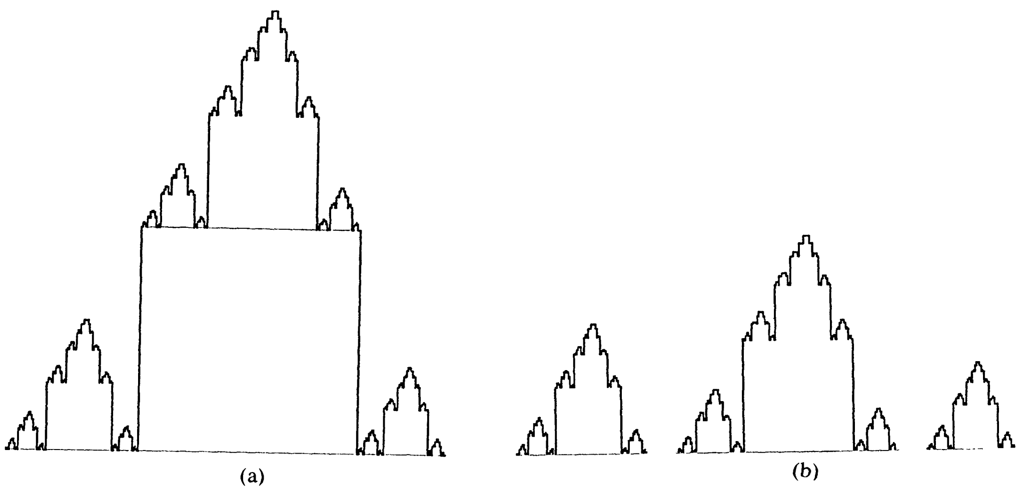


Figure 11. Figure 11(a) is a skyline in which each horizontal line is divided in the proportions of 0.3 : 0.5 : 0.2 with a square constructed on the middle segment. Figure 11(b) consists of 3 parts of Figure 11(a), each of which is similar to the original.

To find A, find the area of the square under the dotted line, which is p_2^2, and the areas of regions 1, 2 and 3, seen in Figure 11(b). Denote the areas of regions 1, 2, and 3 by a_1, a_2, and a_3, respectively. Then

$$A = p_2^2 + a_1 + a_2 + a_3.$$

Region j is similar to Figure 11(a) with scaling factor p_j, $j = 1$, 2, and 3. Thus,

$$a_j = p_j^2 A, \qquad j = 1, 2, 3.$$

Substitution gives

$$A = p_2^2 + \sum_{j=1}^{3} p_j^2 A \qquad \text{or} \qquad A = \frac{p_2^2}{1 - \sum_{j=1}^{3} p_j^2}.$$

FRACTAL DIMENSION. Our knowledge of similarity tells us that the lengths of similar 1-dimensional figures with scaling factor p are related by $p^1 L = l$. Also, when comparing 2- and 3-dimensional similar figures with scaling factor p, their corresponding areas and volumes are related by $p^2 A = a$ and $p^3 V = v$, respectively. More generally, the measures of two similar d-dimensional objects are

related by

$$p^d M = m$$

where p is the scaling factor.

The Koch-like curve of Figure 8 is relatively smooth. As p gets larger, the Koch-like curves get more "curvy", which can be seen in Figures 6 and 9, respectively. At some point does the Koch-like curve become so curvy that it is actually 2-dimensional? To investigate this, let's try to find the measure M (or length if 1-dimensional) of the Koch curve seen in Figure 7(a). In Figure 7(b) is a portion of the Koch curve of Figure 7(a). By actually placing the curve in Figure 7(b) onto four different parts of the Koch curve of Figure 7(a), it is clear that the measures of the curves are related by

$$(1/4)M = m,$$

where m is the measure of the curve in Figure 7(b). The scaling factor relating the two is $p = 1/3$. Therefore, if the Koch curve is 1-dimensional, the measures or lengths should be related by

$$(1/3)M = m.$$

Likewise, if the Koch curve is 2-dimensional, the areas should be related by

$$(1/3)^2 M = m.$$

The relationships $M/4 = m$ and $M/3 = m$ jointly imply that $M = \infty$ or $M = 0$. As a 1-dimensional measure, the length of the Koch curve is infinite, which can be seen by computing the divergent geometric series for the length of the curve. Likewise, the relationships $M/4 = m$ and $M/9 = m$ also imply that the area of the Koch curve is zero or infinity. In this case, the area is zero.

There is a third possible explanation. The Koch curve may have nonzero, finite measure when considered as a d-dimensional object, where $1 < d < 2$. In trying to find the measure of the boundary of the Koch snowflake, we observe that $(1/4)M = m$, but by similarity, $(1/3)^d M = m$. If the measure M is nonzero and finite, then the dimension d must satisfy

$$1/4 = (1/3)^d \qquad \text{or} \qquad d = \frac{\log(1/4)}{\log(1/3)} = 1.26$$

to 2-decimal-place accuracy. This is only a heuristic argument to indicate that fractional dimension may make sense. In fact, this approach can help lead to a workable definition for fractal dimension.

Let's use this procedure to find the dimension of the boundary of the Koch snowflake variation where the initial line is divided in the proportions of $p : q : p$. Consider the Koch-like curve which borders one side of the equilateral triangle. In what dimension d is the measure M of this curve nonzero and finite? To answer this question, observe that this curve is made of 4 identical curves, all of which are similar to the original curve with a scaling factor of p. Note that $p = 0.26$ in Figure 8 while $p = 0.4$ in Figure 9. Denote the measure of each of the four smaller curves by m. Then $(1/4)M = m$ since the large curve is made of 4 copies of the small curve.

Suppose that the Koch-like curve is d-dimensional. Since the scaling factor relating the large curve to its parts is p, then

$$p^d M = m.$$

Since M is nonzero and finite, the dimension d must be such that $1/4 = p^d$, or

$$d = \log 0.25/\log p \quad \text{where} \quad 0.25 \leq p \leq 0.5.$$

Thus, the curve in Figure 8 has dimension $d = 1.03$, while the curve in Figure 9 has dimension $d = 1.51$, both to 2-decimal-place accuracy.

The previous relationship could be written as

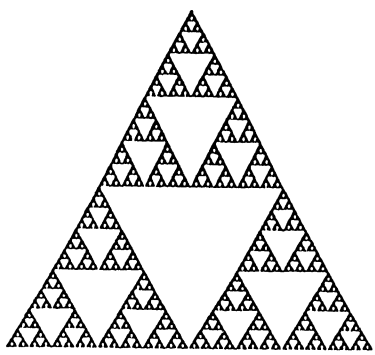
$$p = (0.25)^{(1/d)}.$$

Thus, substituting any dimension d, between 1 and 2, into this formula, we can find a p value that gives us a snowflake variation whose boundary is that dimension. For example, letting $d = 1.5$, we get that $p = 0.397$, to 3-decimal-place accuracy.

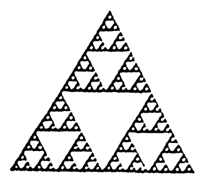
Graphing either $p = (0.25)^{(1/d)}$ or $d = \log 0.25/\log p$ using a graphing calculator or computer graphing program with $0.25 \leq p \leq 0.5$ and $1 \leq d \leq 2$ demonstrates that the dimension d varies continuously from 1 to 2.

This process can easily be used to determine the dimension of other fractal figures. Figure 12(a) is the Sierpinski triangle which is constructed by repeatedly connecting midpoints of all equilateral triangles left and deleting the middle triangle. To find the measure M of what is left, redraw one of the corners, which is seen in Figure 12(b). Denote its measure by m. Since Figure 12(b) can be placed on Figure 12 (a) exactly 3 times, we see that

$$(1/3)M = m.$$



(a)



(b)

Figure 12. Figure 12(a) is the Sierpinski triangle. Figure 12(b) is one corner of the Sierpinski triangle, which is similar to the original.

But the base of Figure 12(b) is half the base of Figure 12(a) so the scaling factor is $p = 0.5$. This means, by similarity, that

$$(1/2)^d M = m.$$

If the measure M is nonzero and finite, then

$$1/3 = (1/2)^d \quad \text{or} \quad d = \frac{\log(1/3)}{\log(1/2)} = 1.585$$

to 3-decimal-place accuracy.

As a last example, let's construct and find the dimension of the generalized Sierpinski carpet. Construct a square with sides of length 1. Divide each side into the proportions of $p:q:p$ where $2p + q = 1$. Keep the squares in the four corners whose sides are of length p. Repeat this process on each of the remaining

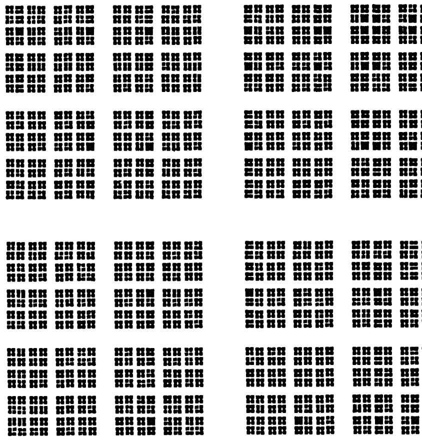


Figure 13. The generalized Sierpinski carpet in which strips that are 10 percent the width of the square are removed from the middle of each square, both horizontally and vertically.

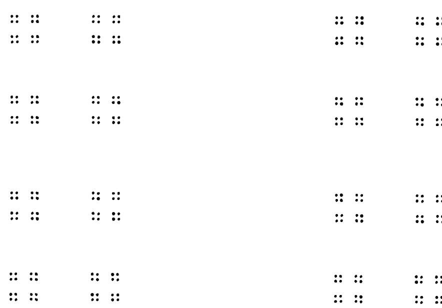


Figure 14. The generalized Sierpinski carpet in which strips that are 50 percent the width of the square are removed from the middle of each square, both horizontally and vertically.

four squares. The limit to this process is the generalized Sierpinski carpet. Figure 13 is the generalized Sierpinski carpet with $p = 0.45$ while Figure 14 is the generalized Sierpinski carpet with $p = 0.25$. The term, Sierpinski carpet, usually refers to the case in which $p = 1/3$.

Denote the measure of the generalized Sierpinski carpet by M. Note that the generalized Sierpinski carpet is composed of 4 equal parts, each of whose measure we will call m. Thus,

$$(1/4)M = m.$$

Also, each of these parts is similar to the whole, with scaling factor p. Thus, if the region is d-dimensional, then each corner is related to the whole region by

$$p^d M = m.$$

For the measure M to be nonzero and finite, then

$$p^d = 0.25 \quad \text{or} \quad p = (0.25)^{(1/d)} \quad \text{or} \quad d = \frac{\log 0.25}{\log p}.$$

In Figure 13 in which the scaling factor is $p = 0.45$, the generalized Sierpinski carpet has dimension $d = 1.74$. Graphing either the function $p = (0.25)^{(1/d)}$ or the

function $d = \log 0.25/\log p$ shows that the dimension varies continuously from 0 to 2 as the scaling factor p varies from 0 to 0.5. In particular, letting $d = 1$ gives that $p = 0.25$, which is seen in Figure 14.

Does it make sense that Figure 14 is 1-dimensional? A heuristic argument that it does make sense is the following. The first step to constructing Figure 14 leaves four squares, with sides of length $1/4$. The upper 2 squares can be moved into the gap between the lower 2 squares, giving a rectangle with sides of length 1 and $1/4$. The second step leaves 16 squares with sides of length $1/16$. Note that these 16 squares can be rearranged similarly to form a rectangle with sides of length 1 and $1/16$. Thus, after n steps, the 4^n squares can be rearranged to form a rectangle with one side of length 1 and the other of length $(1/4)^n$. Thus, the rectangles approach a line of length 1.

The reason we did not include a picture of a generalized Sierpinski carpet with dimension less than 1 is that the resulting figure is much the same as Figure 13, except that you don't see as much because the "squares" are too small.

This approach can also be used to find that the dimension of the Cantor set, in which the middle q is removed from each line segment, is

$$d = \log 0.5/\log p.$$

Note that this is half of the dimension of the generalized Sierpinski carpet. This makes sense since the generalized Sierpinski carpet is the Cartesian product of two Cantor sets.

ACKNOWLEDGMENTS. I used MicroWorlds software for making all of the figures in this paper and would like to thank Logo Computer Systems Inc. for supplying me with a copy of this software. This software is used in many elementary schools. It is also possible to program the TI-82 to draw spirals inside of squares, similar to Figure 2. I would be glad to send my MicroWorlds procedures and/or my TI-82 program to interested readers. I would like to thank Bruce Torrence who made several of the high quality fractal figures for me using Mathematica. These figures had slightly better resolution than the MicroWorlds figures. I would also like to thank Jonathan Choate for his suggestion of the variations of the snowflake curve, and Ray Bobo and Rosalie Dance, who made numerous helpful suggestions after reading early drafts of this paper.

REFERENCES

1. Peak, David, and Michael Frame, *Chaos Under Control: The Art and Science of Complexity*, Freeman, New York, 1994.
2. Peitgen, H.-O., Hartmut Jurgens, and Dietmar Saupe, *Fractals for the Classroom: Part One, Introduction to Fractals and Chaos*, Springer, New York, 1992.
3. Sandefur, James T., *Discrete Dynamical Modeling*, Oxford University Press, New York, 1993.

Mathematics Department
Georgetown University
Washington, DC 20057
sandefur@guvax.georgetown.edu

The Euler Characteristic and Pólya's Dream

Peter Hilton and Jean Pedersen

Dedicated to the memory of George and Stella Pólya

1. INTRODUCTION. The authors were very privileged to have been on the friendliest terms, during the later part of his long life, with the outstanding mathematician and teacher George Pólya—and with his equally remarkable wife Stella. Pólya would often muse out loud about the inspired work of the great mathematicians of the past and wonder how they got their brilliant ideas, what was the source of their profound insights.

Once, in such speculative mood, Pólya was discussing the astonishing observation of Euler (see [2]) that, if one takes any convex polyhedron in the most simple, geometric-combinatorial sense, and counts the number of vertices V, the number of edges E, and the number of faces F, then

$$V - E + F = 2. \tag{1.1}$$

The nature of the formula (1.1) indicates that Euler was thinking of the polyhedron as 2-dimensional, and indeed, in modern terminology, as homeomorphic to the (2-dimensional) sphere S^2.

Pólya speculated that Euler would have been guided by his knowledge of the 1-dimensional case; in other words, of polygons (in the sense of closed polygonal paths) homeomorphic to the circle S^1. He imagined Euler saying to himself: We regard two polygons as equivalent if the vertices of one can be matched with the vertices of the other, and the edges of one with the edges of the other so that the incidence relations *vertex v belongs to edge e* are preserved. Then two polygons are equivalent if and only if they have the same number of vertices. Now it is easy to generalize the notion of equivalence—we regard two polyhedra as equivalent if the vertices, edges and faces of one can be matched with the vertices, edges and faces, respectively, of the other so that the incidence relations *vertex v belongs to edge e, edge e is a side of face f* are preserved. How then are polyhedra to be classified? Pólya dreamed that Euler might have tried to find a similar criterion of equivalence to that which is valid for polygons. In his search, however, Euler in fact discovered the very opposite of what he was looking for—instead of finding what distinguished one polyhedron from another, he found what they all had in common, namely (1.1). In modern terms we say that the Euler characteristic of all such polyhedra is 2.

We elaborate on Pólya's dream in Section 2; we also refer there briefly to the relation of the Euler characteristic to the Descartes *total angular defect*. Then in Section 3 we show how Pólya's dream about Euler's aspirations has, in a sense,

come true. More precisely, by enlarging the concept of polyhedron to include all closed surfaces, the Euler characteristic takes on the role assigned to it in the dream—it does succeed, most admirably, in distinguishing between topologically distinct surfaces.[1]

Of course, generalization doesn't occur in mathematics just to make dreams come true. The concept of polyhedron certainly does not reach its fullest scope with any 2-dimensional configuration; and so we give, in Section 4, an idea of how the concept of polyhedron has grown in the hands of modern topologists, and how the Euler characteristic has itself been adapted to fit into this broader concept. Here one must mention the name of the great French mathematician Henri Poincaré (1854–1912), a pioneer of modern topology, who understood the role of the ***Betti numbers*** of a polyhedron, in connection with solutions of systems of differential equations, and their topological invariance. The more refined version of the Euler characteristic, suitable for any polyhedron of any dimension, is referred to as the Euler-Poincaré characteristic, to mark Poincaré's contribution.

2. PÓLYA'S DREAM. We begin by quoting from [7] where Pólya, in trying to show how Euler might have discovered his famous formula (1.1), was "telling it as he would like to believe it happened." At a certain point we deviate slightly from his actual account in detail (though not in spirit) in order to present his story with what we believe to be the fewest polyhedral examples possible to achieve the desired end.[2] Finally, we describe the Descartes total angular defect Δ of a convex polyhedron. Descartes used spherical trigonometry to prove that $\Delta = 4\pi$ for any such polyhedron. Without giving a proof of this or of (1.1), we give a modified version of Pólya's proof that Euler's formula for polyhedra and Descartes' theorem are equivalent. This remarkable result appears in [9], but, in fact, it came to us in the form of lecture notes, taken by our late friend Dave Logothetti when he attended a talk Pólya gave at Stanford in March, 1974. Thus it is reasonable to argue that, in a sense, Descartes already knew the Euler formula.

Now let Pólya speak for himself. He wrote (see [7]):

> In the "commentatio" (Note presented to the Russian Academy) in which his theorem on polyhedra (on the number of faces, edges and vertices) was first published Euler gives no proof. In place of proof, he offers an inductive argument: He verifies the relation in a variety of special cases. There is little doubt that he also discovered the theorem, as many of his other results, inductively. Yet he does not give a direct indication of how he was led to his theorem, of how he "guessed" it, whereas in some other cases he offers suggestive hints about the ways and motives of his inductive considerations.
>
> How was Euler led to his theorem on polyhedra? I think that it is not futile to speculate on this question although, of course, we cannot expect a conclusive answer. ... One can imagine various approaches to the discovery (rediscovery) of Euler's theorem. I have presented two different approaches

[1] One may remark (with tongue in cheek!) that we have here a beautiful example of the Marxist dialectic in action. We want to distinguish polyhedra, but the Euler characteristic *negates* this. We enlarge the concept of polyhedron and *negate the negation*. Marxism seems to be dead as a political philosophy—but might it be rising phoenix-like from the ashes?

[2] Pólya used six polyhedra and we use five. Can any reader find a suitable sequence with fewer than five polyhedra?

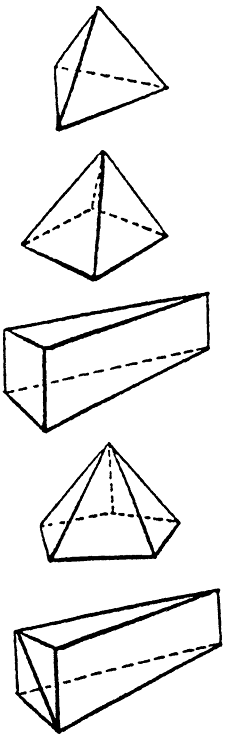
on former occasions.[3] I offer here a third one which, I like to think, could have been Euler's own approach.

1. Analogy suggests a problem. There is a certain analogy between plane geometry and solid geometry which may appear plausible even to a beginner. A circle in the plane is analogous to a sphere in space; the area enclosed by a curve is analogous to the volume enclosed by a surface in space; polygons enclosed by straight sides in the plane are analogous to polyhedra enclosed by plane faces in space.

Yet there is a difference. If we look closer the geometry of the plane appears as simpler and easier whereas that of space appears as more intricate and more difficult. We have a simple classification of polygons according to the number of their sides. ...

So, dreamt Pólya, Euler might have sought to classify polyhedra. First, he might have counted faces.

Here we deviate from Pólya by giving Figure 1, using the same polyhedra as Pólya in the first four cases, but our own fifth polyhedron is simply Pólya's third,



	V	E	F
(a)	4_2	6_3	4_1
(b)	5_2	8_3	5_1
(c)	6_2	9_3	5_1
(d)	6_2	10_3	6_2
(e)	6_2	10_3	6_2

Figure 1.

[3]See [**8**], vol. 1, pp. 35–43 and [**9**] vol. 2, pp. 149–156, and also the annexed problems and solutions in both books. There is also a stimulating discussion of Euler's formula and its proof in [**6**].

the triangular prism, with one of its vertical faces dissected into two triangular faces. One should think of Figure 1 appearing initially without the entries of V and E along the top and without any of the values of V, E and F which are shown.[4]

Now we try to see if the number of faces would classify the polyhedra. So beginning at the top of the table we enter F and fill in the values until we reach an unsatisfactory result (these values have been given the subscript 1 because they were the first to be entered). In fact, we quickly see that although polyhedron (b) and polyhedron (c) have the same number of faces they are not equivalent.[5]

Continuing, from [7] again:

> Here emerges a problem: Let us devise a classification of polyhedra that accomplishes something analogous to the simple classification of polygons according to the number of their sides. Yet in the case of polyhedra taking into account just the number of faces is not enough as the example of Figure 1 shows. ...
>
> What should we do to answer this question? Survey as many different forms of polyhedra as we can and count their faces and vertices?

We now go back to the top of the table in Figure 1 and enter V, then we begin to fill in the missing values for F and V (these values have the subscript 2) until we first reach an unsatisfactory result. What we see is that $(V, F) = (6, 6)$ for the polyhedra (d) and (e). But, again, these polyhedra are not equivalent.

What's left besides faces and vertices? Of course—edges.[6] Again we go back to the top of the table in Figure 1 and enter E, then we fill in the missing values for E (these values have the subscript 3). Alas, this does not work either! As Pólya said:

> They agree also in the number of edges, as they have agreed in the number of faces and vertices. And exploring further cases we find invariably: If two polyhedra have the same F and V, they also have the same E. Thus the number of edges contributes nothing to the classification of polyhedra over and above what the faces and vertices have done already. What a disappointment!
>
> Yet there is something else. If the number E of edges is determined by the numbers F and V... then E is a function of F and V. Which function? Is it an increasing function? Does E increase wherever F increases? Does E necessarily increase with V?... Such or similar questions may lead to more examples... and eventually to the guess
>
> $$E = F + V - 2.$$
>
> An unexpected, extremely simple relation, unique of its kind. What a triumph!

---

[4] We prefer to show our table with the 0-dimensional objects (V), 1-dimensional objects (E), and 2-dimensional objects (F) appearing in that order, from left to right along the top. Although neither Euler nor Pólya did it this way, we argue that it prepares the ground for the generalization of the Euler characteristic to higher dimensions, and also that it allows us to formulate Euler's result so that we can see precisely *what* is invariant.

[5] Pólya said they were *morphologically* different, adding "I am intentionally avoiding the standard term which, by the way, did not exist in Euler's time. One of the ugliest outgrowths of the 'new math' was the premature introduction of technical terms."

[6] According to Pólya, "Euler was the first to introduce the concept of the 'edge of a polyhedron' and to give it a name (*acies*). ... Perhaps Euler introduced edges in the hope of a better classification, and we follow his example here."

Pólya was here discussing his dream of how Euler might have discovered his wonderful formula and how it might have been *first* formulated, so that it is not surprising that it doesn't appear in the form of (1.1).

Now let us turn to a seemingly different aspect of polyhedra. Let us begin with a convex polyhedron P, homeomorphic to S^2. Euclid proved that the sum of the face angles at any vertex of P is less than 2π; the difference between this sum and 2π is called the ***angular defect*** at that vertex. If we sum the angular defects over all the vertices of P we obtain the ***total angular defect*** Δ; René Descartes proved that $\Delta = 4\pi$ for every convex polyhedron P. Thus, for example, there are 8 identical vertices on the cube and the angular defect at every vertex is $\pi/2$, so that the total angular defect Δ is 4π.

We now follow Pólya's line of reasoning to show that $V - E + F = 2$ and $\Delta = 4\pi$ are equivalent statements. In the course of doing so, we obtain a result that takes us far beyond the domain of convex polyhedra. In fact, what we will prove is that

$$\Delta = 2\pi(V - E + F), \tag{2.1}$$

for any *closed rectilinear surface*. This immediately establishes the equivalence of $V - E + F = 2$ and $\Delta = 4\pi$ in the case of convex polyhedra homeomorphic to S^2. Let S be the total number of sides of the faces of our surface.[7] Then, since every edge of the surface is a side of exactly two faces,

$$S = 2E. \tag{2.2}$$

Now let P be a polyhedron homeomorphic to S^2 subdivided into V vertices, E edges and F faces, so that every edge is incident with exactly two faces. Number the vertices $1, 2, \dots, V$ and let the sum of the plane face angles at the nth vertex be σ_n. Then the angular defect at the nth vertex is

$$\delta_n = 2\pi - \sigma_n. \tag{2.3}$$

Note that δ_n will be positive if P is convex, but that, in general, δ_n may be negative or zero. Let

$$\Delta = \sum_{n=1}^{V} \delta_n. \tag{2.4}$$

We are now ready to prove (2.1). What we do is count the *sum of all the face angles* (which we call A) in two different ways. We first count by vertices. Then

$$A = \sum_{n=1}^{V} \sigma_n = \sum_{n=1}^{V} (2\pi - \delta_n) = 2\pi V - \Delta. \tag{2.5}$$

We next count by faces. Now if a face has m sides, then the sum of the interior angles of that face is $(m - 2)\pi$. Thus, if our polyhedron has F_m m-gons among its faces, those m-gons contribute $(m - 2)F_m\pi$ to the sum of the face angles. We thus arrive at the key formula

$$A = \sum_m (m - 2)F_m\pi = \left(\sum_m mF_m - 2\sum_m F_m\right)\pi. \tag{2.6}$$

Now

$$F = \sum_m F_m. \tag{2.7}$$

[7] It is very important to the understanding of this proof to distinguish between the meaning of a *side* and an *edge*. Unfortunately, this is often not done. It is even common to confuse 'side' and 'face'.

Also, since each m-gon has m sides, the contribution to the number of sides from the m-gonal faces is mF_m. Thus

$$S = \sum_m mF_m. \tag{2.8}$$

We put together (2.6), (2.7) and (2.8) to infer that

$$A = (S - 2F)\pi. \tag{2.9}$$

Comparing (2.9) with (2.5) we conclude that $2\pi V - \Delta = (S - 2F)\pi$, or that

$$\Delta = \pi(2V - S + 2F). \tag{2.10}$$

But, of course, if $S = 2E$, we obtain (2.1).

Remarks. (i) Pólya obtained (2.1) without introducing either S or A. However, by introducing these terms, we arrive at the more general equation (2.10). We thereby see that Pólya's argument immediately generalizes, not only to arbitrary 2-dimensional closed surfaces, which need not even be orientable (for details see [**4**]), but even to arbitrary 2-dimensional polyhedra in the most general sense.

(ii) Grünbaum and Shephard proposed the excellent idea of a dual for Descartes' theorem for polyhedra in [**3**]. The introduction of S along with its dual R (for the number of rays) and (2.10) resulted in a perfectly general formulation of that duality (see [**5**] for details). In fact, while $\Delta = 2\pi\chi$ is only true for closed surfaces, $\Delta' = 2\pi\chi$ (where $\Delta' = \pi(2F - R + 2V)$) is true for *any* 2-dimensional polyhedron in the most general sense.

(iii) René Descartes (1596–1650) and Leonhard Euler (1707–1783) worked on these subjects independently—yet, as we have seen, George Pólya (1887–1985) has shown, in an elementary fashion, that their seemingly different, and profound, formulae for convex polyhedra homeomorphic to S^2 are entirely equivalent to each other. It is a trivial theorem that Descartes did not know about Euler's work. It is a less obvious theorem that Euler could not have known about Descartes' work—since Descartes' work on this matter [**1**] was not printed until a century after Euler's death.

(iv) The problem which Pólya believed Euler set out to solve remains unsolved to this day! Nevertheless...

3. ...THE DREAM COMES TRUE. So Euler could not classify polyhedra by the simple method Pólya envisaged in his dream; instead he found the property, which we have already mentioned,

$$V - E + F = 2, \tag{3.1}$$

common to *all* polyhedra homeomorphic to S^2. Is that the end of the story? Certainly not! For topologists have found it useful—indeed, necessary—to enlarge the concept of polyhedron, and then the quantity $V - E + F$, which we call the *Euler characteristic*, does serve to distinguish between types of polyhedra.

The first generalization we will adopt is, as you would expect, that of a *closed orientable surface*. Thus we now consider topological spaces[8] S such that each point p of S has a neighborhood homeomorphic to an open disk. The condition of orientability is equivalent to the requirement that S should be embeddable in $\mathbb{R}^3$. We may always realize S by a rectilinear model, so that S consists of vertices,

[8] S no longer designates the number of sides!

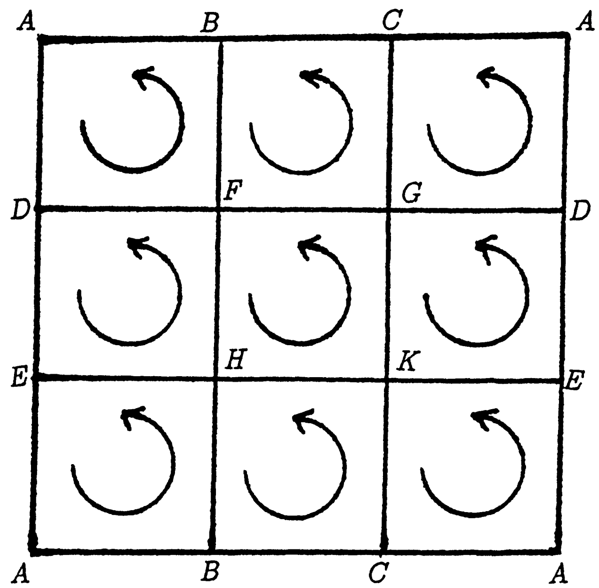


Figure 2.

edges and faces. Then the condition of orientability asserts that each face may be oriented so that a common side of two faces receives *opposite* orientations from those two faces. A typical, and important, example of a closed orientable surface is the torus; a rectilinear model of the torus, with its faces coherently oriented, is given in Figure 2.

By the use of homology theory, one may show that the Euler characteristic χ is a topological invariant; that is, if two surfaces S_1 and S_2 are homeomorphic, then $\chi(S_1) = \chi(S_2)$. This is really very remarkable, since $\chi(S)$ is defined combinatorially, using a subdivision of S into faces, edges and vertices. Nevertheless $\chi(S)$ depends only on the underlying topology of S and not on the combinatorial structure imposed on S.

This result explains why (3.1) holds for all polyhedra homeomorphic to S^2. For such spaces would all have had the same Euler characteristic. That $\chi = 2$ for such spaces follows from the proof of the topological invariance of χ. For the basic theorem is that

$$\chi = p_0 - p_1 + p_2, \tag{3.2}$$

where p_i is the ith *Betti number* of the space. Let us give an indication of what that means. With any topological space K—but let us think of a (finite) rectilinear complex for simplicity—we may associate certain abelian groups H_rK, $r = 0, 1, 2, \ldots,$ called the *homology groups* of K, which, roughly speaking, count the r-dimensional 'holes' in K. If the space is n-dimensional we only have homology groups up to dimension n. Then p_r is the *rank* of H_rK. Now for a closed orientable surface S_g of *genus* g, that is, with g holes or g handles (see Figure 3), the Betti numbers are given by $p_0 = 1$, $p_1 = 2g$, $p_2 = 1$, so that

$$\chi(S_g) = 2 - 2g. \tag{3.3}$$

Of course, the sphere S^2 has no holes, $g = 0$, so $\chi(S^2) = 2$, explaining the results of Section 2 embodied in (3.1). We see from (3.3) that $\chi(S)$ can take as values any

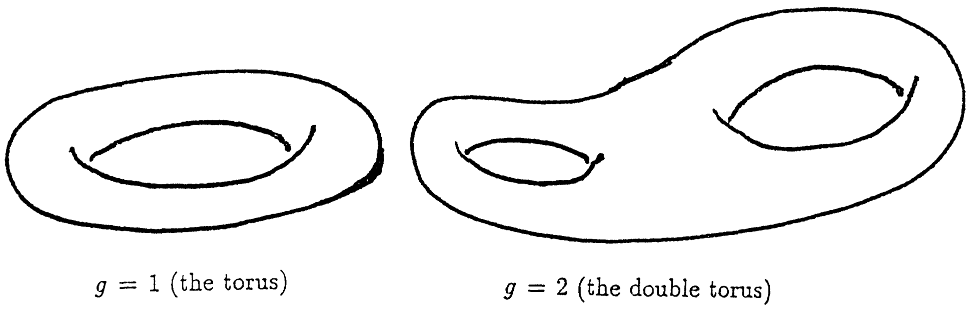


Figure 3.

even number no greater than 2. Thus Euler's aspirations are realized, the dream comes true; $\chi(S)$ *does* distinguish between various closed orientable surfaces. It takes different values on non-homeomorphic surfaces, and is, in fact, entirely determined by the genus of S, according to formula (3.3).

Can we generalize further? Well, we can drop the requirement of orientability. We then obtain a family of closed surfaces characterized by the number k of cross-caps (Möbius bands) inserted into the sphere. The case $k = 1$ is the best known; we then have the real projective plane $\mathbb{R}P^2$ (see Figure 4).

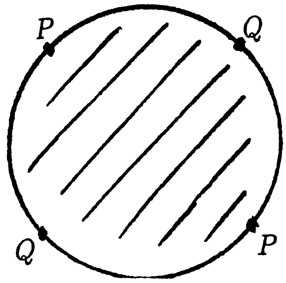


Figure 4. $\mathbb{R}P^2$, represented as a circular disk with diametrically opposite points on the boundary identified.

Now formula (3.2) holds for any compact 2-dimensional polyhedron, in the broadest possible sense. For a non-orientable surface $S^{(k)}$ with k cross-caps, we have $p_0 = 1$, $p_1 = k - 1$, $p_2 = 0$, so that

$$\chi(S^{(k)}) = 2 - k; \tag{3.4}$$

in particular

$$\chi(\mathbb{R}P^2) = 1. \tag{3.5}$$

From (3.4) we see that we can realize any integer no greater than 2, even or odd, as the Euler characteristic of a closed surface, by allowing non-orientable surfaces. However, we now lose the capacity of χ to specify topological type, since an orientable surface of genus g and a non-orientable surface with $2g$ cross-caps have the same Euler characteristic.

Can we achieve any integer as the Euler characteristic of a 2-dimensional polyhedron? The answer is certainly yes, provided that we further broaden the concept of polyhedron. We have already slipped in a reference to a (finite)

rectilinear complex K, that is, a space broken up into vertices, edges and faces, where a (closed) face is simply a polygonal region. Then it is plain that, for any positive integer $n \geq 2$, we can construct a 2-dimensional polyhedron K with $\chi(K) = n$ by taking K to be a bunch of $(n - 1)$ balloons (see Figure 5). For then $p_0 = 1$, $p_1 = 0$, $p_2 = n - 1$.

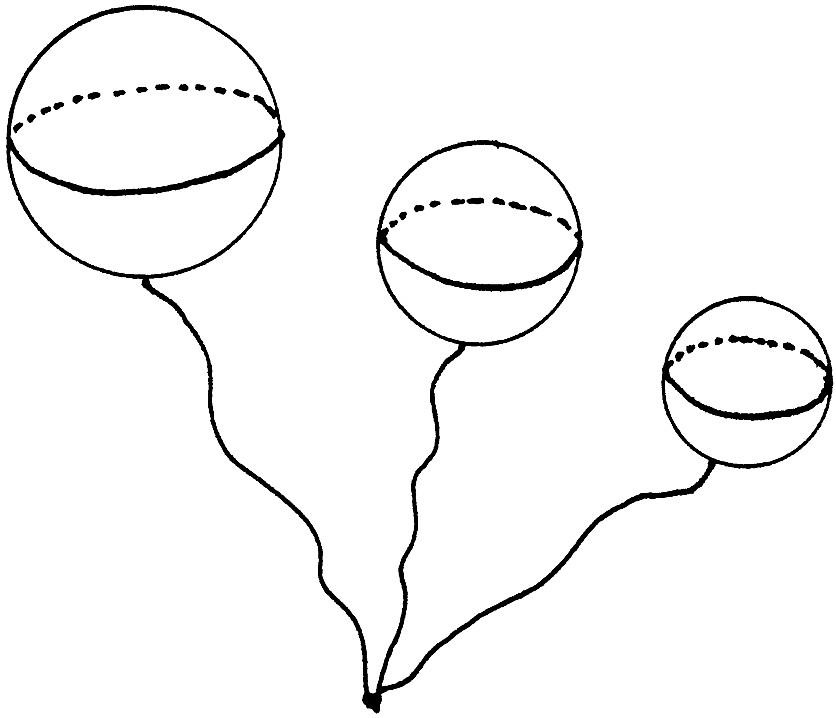


Figure 5. A bunch of n balloons ($n = 3$).

4. FURTHER GENERALIZATIONS. The obvious generalization of a closed orientable surface is that obtained by dropping the requirement that our space be 2-dimensional. We are thus led to the concept of a *closed orientable manifold* of dimension n. Just as S^2 was the first case of a closed orientable surface to be studied, so is S^n the natural starter for a study of closed orientable n-manifolds. Indeed, Schläfli **[10]** generalized (3.1) to show, in effect, that

$$\chi(S^n) = \begin{cases} 2, & n \text{ even} \\ 0, & n \text{ odd} \end{cases} \tag{4.1}$$

Here we regard χ as the alternating sum

$$\chi = \sum_{r=0}^{n} (-1)^r \alpha_r, \tag{4.2}$$

where α_r is the number of r-dimensional cells in some cellular decomposition of the manifold. Thus (4.1) generalizes (3.1); the corresponding generalization of (3.2), which proves the topological invariance of χ, is

$$\chi = \sum_{r=0}^{n} (-1)^r p_r, \tag{4.3}$$

where p_r is the rth Betti number, as before. It is customary to call χ, so generalized, the *Euler-Poincaré characteristic*, since Poincaré proved the topological invariance of the Betti numbers. (It appears that we owe to Emmy Noether the very important observation that we should be dealing with certain *abelian groups*, namely the homology groups, rather than with *numbers*, which are merely their ranks. The homology groups are the true topological invariants.) The proof of the equivalence of (4.2) and (4.3) is a nice exercise in linear algebra.

Of course, (4.1) follows immediately from (4.3) since, for the sphere S^n,

$$\begin{aligned} p_0 &= p_n = 1; \\ p_r &= 0, \quad r \neq 0, n; \end{aligned} \tag{4.4}$$

however, Schläfli used basically combinatorial arguments in [10].

We observe that the equivalence of (4.2) and (4.3) is valid for *any* n-dimensional polyhedron in the most general sense. We may, if we like, confine attention to those spaces admitting the structure of an n-dimensional simplicial complex K. Such a complex K is a collection of *simplexes* of dimension $0, 1, \ldots, n$, where an r-simplex is the convex hull of a set of $(r + 1)$ independent points. Thus

a 0-simplex is a vertex,
a 1-simplex is an edge,
a 2-simplex is a triangle,
a 3-simplex is a tetrahedron, ...

Moreover, distinct simplexes intersect (if at all) in a common face of both. Since the Euler-Poincaré characteristic is a topological invariant, we may, of course, define it for any space homeomorphic to the underlying space of a finite simplicial complex—thus, in particular, for a genuine, geometric n-sphere. If we define a (compact) polyhedron as such a homeomorph, we have achieved a very substantial generalization, taking us as far as we would wish to go in this direction in this article.

Let us point to just one obvious advantage of having generalized the Euler characteristic as we have done. Given two spaces X and Y we can form their topological product $X \times Y$. Now if X, Y are compact polyhedra one may show that $X \times Y$ is also a compact polyhedron. It is then natural to ask for the connection between $\chi(X)$, $\chi(Y)$ and $\chi(X \times Y)$. The answer is delightfully simple!

Theorem 4.1. $\chi(X \times Y) = \chi(X)\chi(Y)$.

The proof is no less pleasing. As one seeks to calculate the homology groups of $X \times Y$ in terms of those of X and Y, one finds a very simple formula if one confines oneself to the case of homology groups with rational coefficients (or, indeed, with any *field* of coefficients), namely,

$$H_r(X \times Y; \mathbb{Q}) = \bigoplus_{s+t=r} H_s(X; \mathbb{Q}) \otimes H_t(Y; \mathbb{Q}) \tag{4.5}$$

(recall that these homology groups are, in fact, vector spaces over $\mathbb{Q}$). From (4.5) we immediately infer that

$$p_r(X \times Y) = \sum_{s+t=r} p_s(X)p_t(Y). \tag{4.6}$$

There is a very nice way to express (4.6). Let us consider the formal polynomial

$\Sigma_{r \geq 0} p_r(X) x^r$. We call this the *Poincaré polynomial* of X, and write it $P_X(x)$. Then (4.6) simply asserts that

$$P_{X \times Y}(x) = P_X(x) P_Y(x). \tag{4.7}$$

This almost completes the proof of Theorem 4.1—and really explains why it is true. For plainly

$$\chi(X) = P_X(-1). \tag{4.8}$$

Theorem 4.1 could not have been brought into existence without the generalization of the Euler characteristic; moreover, it gives us great confidence that we've chosen the *right* generalization. Finally, our proof, like any good proof, leads us smoothly into further questions which, unfortunately, we cannot allow ourselves to be specific about here.

We have not attempted in this article to bring the story right up to date—research continues today on modified versions and adaptations of the Euler-Poincaré characteristic. The progress seems, in some sense, to be inevitable—there is always, somewhere, a George Pólya having a dream!

REFERENCES

1. Descartes, René, *Oeuvres*, vol. X, pp. 265–269.
2. Euler, Leonhard, *Opera Omnia*, series i, vol. 26, Orell Füssli Verlag, 1953.
3. Grünbaum, Branko, and Geoffrey Shephard, A dual for Descartes theorem on polyhedra, *Math. Gaz.* 71 (1987), 214–216.
4. Hilton, Peter, and Jean Pedersen, Descartes, Euler, Poincaré, Pólya—and polyhedra, *L'Enseignement Mathématique, T.* XXVII, No. 3–4 (1981), 327–343.
5. Hilton, Peter, and Jean Pedersen, Duality and Descartes' angular deficiency, *Computers Math. Appl.* 17, No. 1–3, (1989), 73–88.
6. Lakatos, Imre, *Proof and Refutations: The Logic of Mathematical Discovery.* Cambridge University Press, 1976.
7. Pólya, George, Guessing and Proving, *Two-Year College Mathematics Journal* 9, No. 1 (1978), 1–9.
8. Pólya, George, *Mathematics and Plausible Reasoning*, Princeton University Press, 1954.
9. Pólya, George, *Mathematical Discovery*, Combined Edition, John Wiley and Sons, Inc., 1981.
10. Schläfli, Ludwig, Theorie der vielfachen Kontinuität, *Denkschriften der Schweizerischen naturforschenden Gesellschaft* 38 (1901), 1–237.

Department of Mathematical Sciences
SUNY Binghamton
Binghamton, NY 13902-6000

and

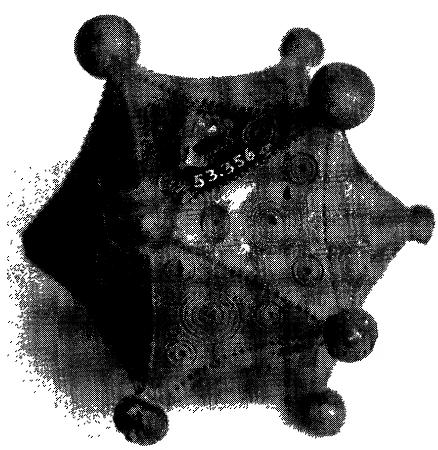
Department of Mathematics
University of Central Florida
Orlando, FL 32816-6990

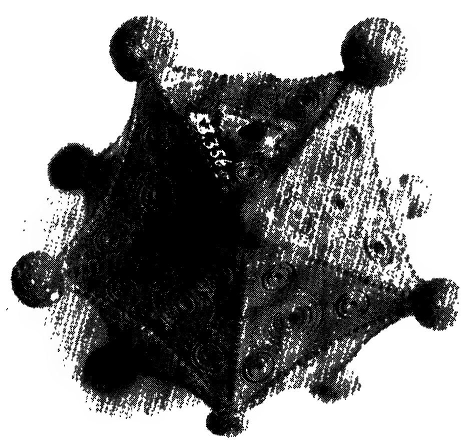
Department of Mathematics
Santa Clara University
Santa Clara, CA 95053

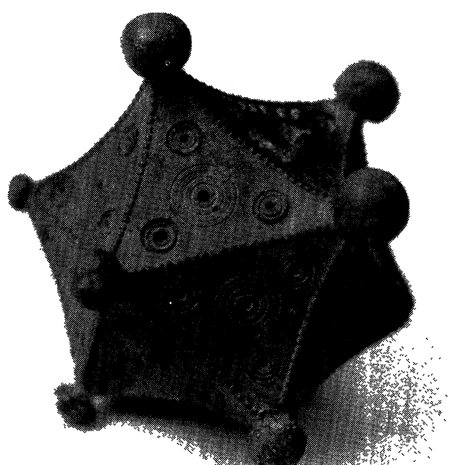
A Roman Icosahedron Discovered

Benno Artmann

During the time from about A.D. 100–300, dodecahedra were apparently popular in the Roman provinces along the Rhine river and in what later became Switzerland, France and England. I have reported about these dodecahedra in [1]. A detailed account from the archeological point of view has been given by R. Nouwen [3] (in Dutch).







Together with 77 dodecahedra, Nouwen shows one singular icosahedron on p. 59. Like the dodecahedra, it is hollow and made from bronze. It weighs 465 gr and its overall diameter is about 8 cm.

The icosahedron was excavated in 1953 in the village of Arloff (some 30 km SW of Bonn). At that time, it was erroneously classified as a dodecahedron and put into storage in the Rheinisches Landesmuseum in Bonn. There it sat in the basement until recently, when Dr. Ursula Heimberg on the staff of the museum had another look and discovered that it was not at all a dodecahedron, but an icosahedron. (Its museum number is Rheinisches Landesmuseum Bonn Inv. 53.356.) As Nouwen says, this icosahedron is even more mysterious than the dodecahedra. Nobody knows what its use or purpose was.

From the time of Plato, dodecahedra have been associated with the universe or the zodiac. There are natural crystals of pyrite in the form of (almost) regular dodecahedra. That is, dodecahedra may come from a context outside of mathematics and get their meaning from there.

Icosahedra seem to be different. They were discovered "inside" mathematics and have their specific significance only as members of the class of regular polyhedra. (For details compare Waterhouse [4] or my [2].) It is hard to imagine anything other than a mathematical origin of the icosahedral form of the object from Arloff. That is, at least a little knowledge of Euclid's "Elements" must have diffused to the northwestern provincial regions of the Roman Empire. Mathematical knowledge must have been much more widespread during imperial Roman times than what was supposed until now.

To say more would be pure speculation. May we conclude that it was the logo of the GRMA (Gallo-Roman Math. Association), or even of the DMA (Druidic Math. Assoc.)?

REFERENCES

1. B. Artmann, Roman Dodecahedra. *The Math. Intelligencer* **15**, no. 2 (1993), 52–53.
2. B. Artmann, Symmetry through the ages. In: J. M. Anthony, ed.: In Eves' Circles, 139–148, MAA 1994.
3. Nouwen, R., DeGallo-Romeinse Pentagon-Dodecaeder: Mythe en Enigma. Publications of the Gallo-Romains Museum Tongeren (Belgium) Nr. 45. Hasselt 1993.
4. Waterhouse, W., The discovery of the regular solids. *Arch. Hist. Exact Sci.* 9 (1972–1973), 212–221.

Fachbereich Mathematik
Technische Hochschule Darmstadt
64289 Darmstadt
Schloßgartenstr. 7
Germany

A Fable of Reform

John J. Schommer

For years the math department at Rolling Hills State University (RHSU) had been taking heat from nearly every liberal arts department on campus for being too tough on liberal arts students. Having refused on many occasions to adopt the "Math for Poets" class that many universities offer to fulfill university math requirements, the math department at RHSU had been adamant in its support of College Algebra as the *least* course a student could take to fulfill this requirement. But reform was in the air...

1. THE DECISION. Maybe it was the fact that a lot of new folks had just been hired by the math department. Maybe a lot of folks in the math department just really had a good breakfast that morning. For whatever reason, someone suggested at the faculty meeting one afternoon that perhaps the department should offer a course in chess. At first the surprisingly enthusiastic reception was aimed at adopting the course as an elective. The department soon found itself considering seriously, however, whether in fact chess should be adopted as a way of fulfilling the university math requirement. Enthusiasm for the idea, though far from unanimous, was in the majority that afternoon, and a committee was soon formed with many eager volunteers to draw up a syllabus for the course, as well as to lay out the best arguments for a chess course before the university curriculum committee.

The newly formed chess committee had barely met for an hour when they realized that, from a professional standpoint, most of what they wanted to accomplish in a chess class would resonate strongly with the NCTM Professional Standards for Teaching Mathematics [**1**, pp. 19–67]. Within a week of their first committee meeting, a very eager chess committee had completed the first draft of its report to the curriculum committee. It emphasized the following:

- **Critical Thinking / Problem Solving.** If ever there was a game that involved critical thinking it was chess. Every time the board changed new problems were created, problems with many valid game-winning solutions, problems completely unlike the pre-fabricated exercises of the typical algebra book with their stock unique solutions. With the right teacher at the helm of this class, each lesson might very well involve almost 50 minutes of careful, critical analysis.
- **Collaborative Learning.** A well-run chess class could very well serve as the model for a collaborative learning environment. Again, with the right teacher at the helm, a class could easily be conducted in small groups, all members tossing out ideas, discussing the critical pros and cons of every move on the board, moving only after achieving some democratic consensus for a course of

action. No sleepy lecture sessions, classrooms would be noisy and bubbling with intellectual ferment. Quiet shy students with little self-esteem would find their views solicited and even adopted on occasion. Students, even shy students, would learn to make their best arguments before a group of their peers. Highly disciplined and yet fun, chess promised to be an almost perfect venue for collaborative learning.

- **No "mindless" symbol manipulation.** College algebra's harshest critics were almost incapable of saying "symbol manipulation" without saying "mindless" at the same time. If chess were adopted as a course, the math department could begin to demolish that ugly characterization of their discipline. Aside from the fairly easy-to-learn notation system for recording games (a skill that would be required only rarely in this class), 99% of the time nothing vaguely resembling abstract symbols would be used! Abstract symbol manipulation would be dead, the demand for critical analysis would be maintained. The best of both worlds!
- **Physical Manipulatives.** Here was something that you rarely saw in a college math class—physical manipulatives! And beautiful, artistic manipulatives at that. Though some students would be content with their generic plastic chess sets, some would no doubt fall in love with the artistry of more beautiful sets, a love which would surely make them want to play often. The more artistic students might even be motivated to sculpt their own sets. With chess sets becoming valued artistic possessions, could the desire for the beauty of the game itself (that reservoir of critical thinking) be far away?
- **Negligible costs, even for a technology-friendly class.** Though a few students would indeed want fancy chess sets, and some would want to buy chess software for their computers, none of these would be required. The simplest of boards is all that would be needed; a college student could effectively walk into any local toy store and come away with the only materials needed for the course. Furthermore any chess program, a site-license for which would be cheaper than the heavily promoted math software found in most professional math journals, would instantly convert any computer lab into a chess lab.

2. THE PLAN. To say that the university curriculum committee was stunned to hear the math department's proposal would be an understatement. It all sounded so *progressive*. Liberal arts members of the curriculum committee began to imagine life in harmony with the math department, their students no longer complaining of the drudgery associated with completing their math requirements. So caught up in the idea of this major reform in the math department (and perhaps just a little fearful that this opportunity might not ever come again), the committee quickly approved the course. Strange as this might seem now in retrospect, no one ever bothered to ask, where was the mathematics?

Every member of the chess committee was eager to be one of the first to teach the new class. Since only few would be able to, the committee decided that each member should draw up a detailed proposal for the course, and that the members with the best proposals would then become the first teachers of the course. Wonderful proposals were made, and when the teachers for the first run of "Math 130: Chess" were finally picked, they all had a certain similarity in their approach.

- Most of the traditional "lectures" would occur in the first week of the course, consisting primarily of the basic moves of each piece and ways to avoid being

checkmated early. Chess aphorisms like "play to the center of the board" and "try to control a middle square" would be discussed.
- Any lectures in subsequent weeks would be enthusiastic and quite short, focusing on problems associated with various opening, middle, and end games. Most of the time though, students would be actually trying to solve various chess problems in small groups. The emphasis in class would be "problem of the day and play, play, play".
- Students would not be required to memorize games or even openings.
- Chess grandmasters would be invited to talk about life in the world of chess and their favorite matches.

With an enthusiastic group of teachers chosen, there was one last piece of business that had to be decided: how would grades be assigned? Here, the teachers thought, was where they could really exploit the beauty of living in the computer age. With chess programs available that provide competition on several levels of difficulty, putting together an objective grading system would be easy: grades would be determined by students' ability to defeat a given chess program a certain fixed percentage of times. A student could earn an "A", for example, by defeating the program at "level five" 7 out of 10 times. The beauty of this system is that there would be clear objective standards that would mean the same each and every year. Gradewise, the course would be essentially teacher independent!

The teachers worked as a group to find good chess problems with interesting solutions. Great software was found (with dazzling visuals that projected nicely on overhead screens) that would play out the day's problem so that everyone could compare their solutions to the computer approach. Four speakers were lined up for the fall, all of whom traveled extensively because of chess and had marvelous stories to tell of games bravely fought and often won. Two speakers even offered to stay after class and play all takers on the quad in simultaneous games.

By summer's end, it appeared that a dynamic, exciting class had been prepared that would really get first-year students deeply involved in problem solving, developing critical thinking skills that would serve them throughout their lives. Well, at least that was the plan...

3. THE IMPLEMENTATION. There was a lot of excitement that first week. Students and teachers alike felt that they were somehow part of a revolution. When teachers described how the class would be run—very short lectures, lots of game playing—students were pleased and eager to get started. Though teachers tried to stress that in fact this "game playing" was going to be serious business, some students (especially those who had failed college algebra before) couldn't help but think that this was going to be an easy way to complete that old math requirement.

By the second week, it was already becoming quite clear who was probably going to be earning "A"'s at semester's end. Knowing that these students would more than likely find each other after class for some challenging play, teachers made sure that this wealth of natural ability was spread around when assigning folks to their small groups. And these natural players were actually quite happy to be the big fish in those small ponds. As the semester progressed teachers noted some minor problems with certain of these "prodigies" lording over the others in their groups, but this indeed only proved to be a minor problem. All in all, group work turned out pretty well in the first two months.

The semester's guest lecturers were, by and large, quite interesting. Though one presenter's "inside" stories about the chess world proved way too obscure, the presentation by one grandmaster was particularly entertaining and enlightening—the image of chess nerd was thoroughly dispelled. The first "chess on the quad" was surprisingly popular, and an alert university relations officer made sure that local media would be around for the second event. The second event was equally popular, and did indeed find its way into local news stories. In fact, one reporter's story evolved into an examination of mathematics reform on the national level and was picked up later by various news services, much to the delight of a university president wanting very much to be perceived as being on the cutting edge of reform.

It's actually not too hard to pin down when things began to deteriorate—it was right around midsemester, two weeks before grades were due. For fear that the more prodigious players might go for their "A" early in the semester and then start sleeping in, it was decided that you could not leave a level of computer play until after you won at that level the required percentage of times. To get a midsemester "A" you had to graduate to level four by midsemester, to get a "B", level three etc. The network in the computer lab was set up so that one could essentially play at any time of day for official scoring purposes, and was furthermore designed so that a teacher could be certain that the person who claimed to be playing was actually the person on the machine. The computer also kept a good log of how long a person was on the machine.

It was this simple log that shouted warning signs long before midsemester. After the initial spurt of logging on at the beginning of the year, only the naturally talented students seemed to be logging on regularly, and only they seemed to be staying on for periods of time vaguely resembling the fabled ratio of two hours homework to one hour of classtime. Despite these early warning signs, the teachers assumed that students preferred playing each other, and would eventually begin logging on regularly once midsemester approached.

And so they did. With two weeks to go before midsemester grades were due, a trickle and then a torrent of folks began to log on. The only problem was that in the course of all their group work, students had managed to develop a completely unrealistic idea of their own particular chess skills. Beating the computer even at the lowest level proved far harder than they imagined. Office hours were well attended the week before midterm with students absolutely panicked about their grades. The teachers discussed the crisis and decided that they were partially responsible for the mess—they were going to have to be quite generous with midsemester grades. An anonymous survey also revealed that students were in fact *not* playing chess all that often outside of class—the computer logs had been a pretty accurate reflection of who was putting in time with "homework". The teachers made clear, though, that the ultimate grading criteria remained the same—by the end of the semester, a 7 out of 10 success rate at level two would be necessary to receive a minimal passing grade. An understanding dean was notified officially by the teachers about what had been going on (he had heard quite a bit already thank you very much), and it was decided that students would be allowed to drop if they wished (even this late in the semester) with a "WP". A discouraging number chose this option.

The computer log that developed in the two weeks prior to midterm made something else abundantly clear—a large number of students were having difficulty just getting past level one. The computer could replay students' games for the

instructors, and many students seemed to be roaming the board aimlessly. Opening play appeared to be especially bad, leaving students with barely defensible positions. Students also seemed to have no sense of when resignation would be preferable to playing out their weak positions. Of course one quick way to get students into more competitive openings with the computer was to have them memorize famous openings like the Queen's Gambit and the Sicilian Defense. It was precisely this kind of rote memorization which everyone wanted to avoid when the course was being planned. Yet it was clear that something like rote memorization was going to have to happen if many students were going to get past level one. The teachers put together a booklet of about ten famous openings and suggested that students might begin enjoying more success against the computer if they memorized a few of them—not that this was required, mind you. A day was put aside for a more traditional lecture, and the pros and cons of the ten openings we discussed.

A few of the talented kids had discovered books on chess weeks earlier and had in fact learned some of the classic openings already. Their impressive names had been dropped by the "prodigies" in small groups, but it wasn't until the big grade scare that the average student felt particularly motivated to memorize them. In any case, soon after students started memorizing openings, the level one barrier fell for many. The computer did not always respond according to the book, but students found that if they persisted with a prepared opening, more often than not they began to enjoy some success. At this point they had, of course, little appreciation for why such openings were yielding success, but they were at least content to be finally winning and relieved to be "passing" the course. Appreciation for those openings could always develop with time. Some students did want to drop the course a few weeks after midsemester when it became clear that they would have to spend quite a bit of time memorizing a chess "vocabulary". They were not altogether happy when a very tired dean did not allow them to withdraw.

With little more than a third of the semester still to go, a "C" was looking more and more possible to many and the assault on level three had begun in earnest. But the grade scare had had a profound effect on classroom temperament. Memorization had not only brought success to those who were having trouble, it had also brought drudgery. Chess class was no longer the "consequence free" class that it was just a month earlier. Success, it seems, had been a double-edged sword. Though in-class time was itself still considered fun, there was nothing particularly fun about the personal discipline required to study chess outside of class. One of the messages students had received implicitly before the grade crisis was that there were no "right" and "wrong" solutions to chess problems, only "different" ones. The success of memorization seemed to destroy that vision of chess—there apparently were "right" and "wrong" approaches to chess problems (depending on level of play) and the discovery/small group method of learning how to make those distinctions came with the risk that the desired grade might not be earned by semester's end. Students no longer entertained false notions of their own abilities. Teachers began to regularly hear the question, "...but what is the *right* move?" The chess books held in reserve at the library were being used more often as students wanted to know more about the "correct" way to play chess. Ironically, the more tiresome Math 130 had become for many, the more success students seemed to be having according to the computer logs.

Now as uncomfortable as the class had become for those who had a reasonable shot at earning a "C" by finals, it had become *very* uncomfortable for those for

whom a "D" was becoming their best hope. The odd thing though, was that the people who so desperately needed to spend more time with chess were not really logging onto the computer as often as might be thought. After initially becoming very involved in the memorization craze, the participation of the struggling students fell off markedly when it became clear that memorization was not going to be a quick fix to their grade predicaments. It turned out there was still plenty of chess to play after those prepared opening moves. Occasionally a student would find a teacher during an office hour and express the frustration that no doubt many were feeling—they "understood" everything in groups, and they "followed" everything said in lectures, but as soon as they tried to play on their own, they met with failure. Challenged by the records of the computer log that claimed that they were not "studying" anywhere near as often as they needed to, their response was to confess that they indeed had not practiced as much as they should, but that their jobs and extracurriculars took up an important chunk of their time.

Perhaps surprisingly, the biggest threat to the future of the chess class did not come from a student, an "arts" faculty member, or an administrator. The teacher who pushed hardest and worked most enthusiastically for the chess class began to wonder openly whether any of this was an improvement over the old college algebra class. The chess grade distribution looked strikingly similar to the one for college algebra, about the same number of people had dropped, and the same old excuses for poor performance were being heard. Most profoundly discouraging though was the fact that the kind of students who struggled in College Algebra primarily because they didn't do their algebra homework, weren't doing their chess homework either. It was almost precisely for this particular group of students that the chess class was put together. Group work, manipulatives, all the things that were really supposed to get this group involved did not seem to have anywhere near the effect on personal study habits as was hoped. Embraced precisely to bring students formerly alienated from math into a deeper involvement with problem-solving, the chess version of "math reform" had yet to evidence the kind of involvement that instructors were looking for.

The semester ended the way many do. Some students hardly lifted a finger and got an "A". Some students, overconfident from their midsemester "A"'s and "B"'s, lost a letter grade. Some students who struggled with a "C" all semester long and who logged many an hour on the computer finally crossed the "B" threshold with a tremendous sense of accomplishment. Yet other "C" students who worked just as long and hard could not quite cross that "B" threshold, finishing the course quite frustrated and vowing never to play again.

4. BACK TO THE DRAWING BOARD. With grades assigned and the campus virtually deserted, the chess committee got together to debrief the teachers and assess the semester. The course evaluations that students had completed in the last week of classes were opened and read aloud. A consensus quickly emerged that students were quite happy with the conduct of daily classes, and in particular the group work. They were disappointed however that group work didn't seem to factor explicitly in their final grade. One student's comment brought considerable laughter—"There was too much chess". But when the laughter died down everybody had a sense that they knew what the student meant. Almost 90% of class time had been spent solving chess problems, an almost maddeningly efficient use of class time. A small but enthusiastic minority claimed that this was the best college course they had ever taken. An equally small but adamant minority claimed that it

had been the worst. Most of the criticism seemed directed at the grading. Though none could take the grading to task for being subjective, students as a whole did not feel that the final grade they were given adequately captured what they "knew". Any restructuring of the course then, would have to focus on grading.

It seemed that the committee would first have to agree about what precisely they were trying to measure. "Chess ability" rolled easily off the tongue, but was perhaps too vague. Unfortunately, further discussion didn't clear this up much. The committee felt itself going in circles, replacing the vague "chess ability" with equally vague-sounding things like "successful contingency planning" and "successful problem resolution". The word "success" seemed to be the only commonality to all this vagueness, and this suggested to the committee that they were on the right track keeping to a grading system based on numbers of wins.

Perhaps a balance could be struck by setting the hurdles lower and then adding a lot more of them: a "D" could be earned by beating level two 20 out of 40 times playing white; a "C" could be earned by winning at the same level 20 out of 40 times playing black; and so forth through level three (a full two levels below what was needed to get an "A" last semester). This would also solve the problem of those few "A" students who had the course pretty much rapped up after midterm—they would have to log at least 52 games more than last semester to get their "A"'s. Students at lower levels would be rewarded to a greater extent for their quantity of play, if less so for their quality.

But maybe the problem with the grading system went deeper. Despite their quaint but honest attempt to establish what they were trying to measure, perhaps emphasizing each student's individual ability to win was somehow fundamentally wrong-headed. Perhaps the clear equation between students' ability to win at a certain level and their actual problem-solving abilities was just an illusion. To be sure, playing to win seemed to be an integral part of chess. But didn't the current grading system engender a hurtful and oppressive hierarchy among the students? A distasteful elitism? Committee discussion naturally turned to "alternative assessment".

Perhaps portfolios could be required of students. Students could be asked to submit a certain number of their "best" games. The computer program that the department had been using in fact was capable of keeping a record of all games using standard chess notation, so it would not be too hard to submit games to an instructor for evaluation. If wins at any level were the only games that qualified for submission, then there would be little for the instructor to do but make sure that the correct number of predetermined wins was achieved. If the "quality" of each win were to be judged, then perhaps the board position of each of the student's prized games could be entered into the chess program, say after the first 10 moves, and the number of moves it took the student to win the game from that point could be compared to the number of moves the computer would take to win the same game. Of course, the very notion of "quality" threatened to bring back the aura of elitism that the committee was trying to eliminate. Furthermore, given that this grading system was more difficult for the instructor to manage than the original "elitist" grading scheme, it was not clear that this alternative system was really worth the extra trouble. If the department was serious about ending elitism, the notion of quality wins in a portfolio would have to be abandoned.

The committee did come to a quick agreement that some kind of writing could be an important component of the portfolio. Students capable of explaining "Kasparov's 23rd move against Karpov in the third game at Sarajevo" were

certainly demonstrating some advanced knowledge of the game—provided of course their analysis was correct. The committee would have to seriously consider portfolios which included term papers purporting to analyze some of the most famous (and perhaps not so famous) chess matches of all time. Unfortunately, if part of the purpose of changing the grading scheme was to make students happier with the course, it was not at all clear that adding an analytical term paper to their responsibilities was going to accomplish this.

Still other candidates for portfolio submissions were discussed. Back when the chess committee was trying to sell the curriculum committee on the value of having physical manipulatives, their enthusiastic rhetoric suggested that some of the more artistic students would be inspired to sculpt their own chess sets. Perhaps sculpture of this sort would constitute a valid entry in a chess portfolio. In the same vein, perhaps history majors would want to contribute papers detailing chess history, perhaps sociology majors would want to write on the contributions of different ethnic groups to chess, etc.

But just when everybody was feeling wonderfully inclusive about what would comprise a portfolio, someone on the committee, to everyone's horror, realized that the point of this course originally was to satisfy a university *math* requirement. The committee had been on the verge of accepting sculptures of chess pieces as reasonable demonstrations of university-level competence in mathematics! How far they had come! They had not even begun to discuss one of their students' most common complaints—that their group work never factored explicitly in their grades. Reawakened to the original purpose of the course, the committee now realized that allowing group work to count for a significant portion of the final grade would confront them with yet another problem. The math department would effectively be saying "because you are good at *group* work in *chess*, you have sufficiently demonstrated *individual* competence in *mathematics*." That was quite a stretch. A good Outward Bound experience or membership on one of the school's sports teams might as well contribute to satisfying the math requirement—any of those experiences usually involved group problem solving.

As long as the grading in the chess course focused on each individual's ability to win games, there seemed to be a mathematical character to the whole enterprise. Once the individual's ability to win games was removed from its preeminence in the grading scheme, the validity of chess as a substitute for math seemed questionable. Questions indeed appeared to be in no short supply. Why did chess ever seem like a valid way to fulfill the math requirement in the first place? Why precisely did the possibility of alternative assessment seem to negate the value of what the committee was trying to accomplish? Were members of the committee simply closed-minded? Did everyone on the committee have some subconscious elitist agenda? And finally what, if anything, did all this say about the validity of alternative assessment strategies in the typical *math* course?

Everyone was now feeling quite frustrated. When this experiment had been but a twinkle in the committee's eye, the campus support for this particular reform was wide. Unless the department diluted its current grading criteria though, setting lower "hurdles", allowing group work to count significantly in the grading, and perhaps adopting certain alternative assessment strategies, support for this kind of reform would wane and the math department would once again become the whipping boy of the campus. In sum, the math department appeared to have accomplished what few could have thought possible: they had taught far less traditional math than ever before, and yet managed to make the same numbers of

people every bit as resentful. Was all this "reform" worth the effort? The committee adjourned, undecided as to whether Math 130 would be offered the following year.

REFERENCE

1. NCTM, Professional Standards for Teaching Mathematics, Reston: NCTM, 1991.

Department of Mathematics & Computer Science
University of Tennessee at Martin
Martin, TN 38238
jschomme@utm.edu

Could you "integrate" this picture?
(see page 148)

18. Yves Nievergelt, "Graphic Differentiation Clarifies Health Care Pricing," *U.M.A.P. Journal*, Vol. 9, No. 1 (Spring 1988), pp. 51–86, reprinted as UMAP Module 678, COMAP, Lexington, MA, 1988.
19. Yves Nievergelt, "Fisher's Effect: Real Growth is not Interest Less Inflation," *The Mathematics Teacher*, Vol. 81, No. 7 (October 1988), pp. 546–547.
20. Yves Nievergelt, "A Graphic Introduction to Functions: The Federal Income-Tax," *U.M.A.P. Journal*, Vol. 10, No. 2 (Summer 1989), reprinted as UMAP Module 689, COMAP, Lexington, MA, 1989.
21. Yves Nievergelt, *Mathematics in Business Administration*, Richard D. Irwin, Homewood, IL, 1989.
22. Donn R. Pescatrice, "The Inelastic Demand for Wagering," *Applied Economics*, Vol. 12, No. 1 (March 1980), pp. 1–10.
23. Michael C. Reed, "Mainstreaming Calculus Reform," *SIAM News*, Vol. 27, No. 7 (August–September 1994), p. 7.
24. James Studnicki, "Regulation by DRGs: Policy Or Perversion?" *Hospital & Health Services Administration*, Vol. 28, No. 1 (1983), pp. 96–110.
25. Michael E. White and William A. Luksetich, "Heroin: Price Elasticity and Enforcement Strategies," *Economic Inquiry*, Vol. 21, No. 4 (1983), pp. 557–564.

Department of Mathematics, MS-32
Eastern Washington University
526 5th Street
Cheney, WA 99004-2431
ynievergelt@ewu.edu

Mathematics, rightly viewed, possesses not only truth, but supreme beauty—a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the trappings of painting or music, yet sublimely pure, and capable of stern perfection such as only the greatest art can show.

—*Russell*

Answer to Picture Puzzle
(p. 142)

Th.-J. Stieltjes [one of the two only known pictures of him] 1856–1894.

On December 31, 1894, Stieltjes passed away in Toulouse; he was just 38 years old. During the academic year 1994–1995, both in the Netherlands and France events were organized for the commemoration of the 100th anniversary of his death.

Submitted by Professor Jean-Baptiste Hiriart-Urruty
Université Paul Sabatier (Toulouse III)
U.E.R.
Mathématiques, Informatique, Gestion
118, route de Narbonne
F 31062 Toulouse Cédex, FRANCE

Is Optimal Pricing a Myth From Business Calculus? Is Business Calculus an Oxymoron?

Yves Nievergelt

Writing in this MONTHLY about "Two-Year Magazine Subscription Rates" [**8**], Professor Underwood Dudley has put on record a finding and a question of relevance to business mathematics:

> The inescapable conclusion is that mathematics does not enter into the rate-setting process. [...] Mathematics, as one respondent pointed out, is irrelevant. [...] How can that be?

Professor Dudley's question is answered in the present essay.

1. MATHEMATICAL SOCIETIES SHOW COUNTEREXAMPLES TO OPTIMAL PRICING. Firstly, publishers may eschew mathematics in setting rates because they may be following their mathematics professors' counterexamples.

Indeed, none of the dues statements for 1995 from the American Mathematical Society (AMS), Mathematical Association of America (MAA), and Society for Industrial and Applied Mathematics (SIAM) offers any discount to members of less than sixty years of age for payments covering more than one year of subscription. Moreover, while these societies send advanced notices requesting early renewal, none offers any discount for early payments either, which corroborates a paraphrase of Professor Dudley's findings:

> Three magazines in the sample (*Bulletin of the AMS*, *American Mathematical Monthly*, and *SIAM Review*) give no discount at all, completely disregarding their readers' ability to make calculations of present values. [Italics substituted.]

Moving from present to future values, observe that the MAA's dues increased from \$122 to \$128 (for all three MAA journals) from the dues notice sent in 1993 to the notice sent in 1994, a relative increase of about 4.92%, which exceeds twice the annual rate of inflation for that period, about 2.3% [**2**], thus causing a real price increase (Fisher's effect [**19**]) of about $(0.049 - 0.023)/1.023 \approx 2.59\%$, which campus representatives must now sell to their librarian.

Furthermore, the discontinuous rate schedules of the AMS and MAA give members financial incentives to understate their annual professional income by imposing large surcharges for small increases in income: the AMS dues jump from \$87 to \$116 as income jumps from \$45,000 to \$45,000.01, corresponding to a marginal income-tax rate of $(116 - 87)/0.01 = 290{,}000\%$, while the MAA dues jump from \$128 to \$159 as income jumps from \$50,000 to \$50,000.01, corresponding to a marginal income-tax rate of $(159 - 128)/0.01 = 310{,}000\%$.

Thus, we the 57,657 mathematicians listed in the 1994–1995 *Combined Membership List* [**1**] appear to have encountered some insurmountable obstacles to optimal pricing and calculations of present values.

Might the same obstacles also have seemed insurmountable to the typical magazine publisher with only one freshman course in business mathematics, and the typical reader having majored in English with only one course in "college algebra" (for readers outside the U.S., the equivalent of "college alphabet" in English)?

2. ECONOMISTS PROVIDE FURTHER COUNTEREXAMPLES TO OPTIMAL PRICING. Secondly, publishers may eschew mathematics in setting rates because they may be apprehensive of their economics professors' counterexamples.

A successful example of optimal pricing might consist of the following elements:

(1) the name of a business firm (private or public, large or small),
(2) data and the determination of the demand and supply curves for that firm's product,
(3) the theoretical calculation of the optimal price, and
(4) the observation that the firm's profit increased to the predicted maximum after the adjustment of the initial price to the calculated optimal price.

Unfortunately, even though I had occasionally taught business mathematics to freshman business majors and to Executive Masters in Business Administration (EMBA) over a decade [**21**], in a decade of informal search I had been unable to identify and document any successful example of optimal pricing.

Worse yet, a famous example exists that demonstrates a spectacular failure of an attempt at optimal pricing: after New York racetracks adjusted their take-out rate (price of betting) to the optimal rate predicted by theory, the total handle (revenues) reportedly went *down*, and economists disagreed on the reasons [**22**].

Might the same reasons that eluded economists also elude the typical magazine publisher with only one freshman course in microeconomics?

3. ECONOMISTS ALREADY ENCOUNTER DIFFICULTIES IN ESTIMATING THE DEMAND. Thirdly, publishers may eschew mathematics in setting rates because they may remember their economics professors' difficulties.

Optimal pricing requires a knowledge, through a combination of measurements and derivations, of the supply and demand functions for the products under consideration. However, before any optimal pricing can occur, the determination of the relevant demand may already present difficulties to specialists. For example, economists do not seem to agree among themselves whether the demand curve for potatoes in Ireland sloped downward or upward (whence it would earn its qualification as a "Giffen good") in the years 1845–1849 [**12, 15**].

If economists cannot agree on the sign of the slope of the demand curve for potatoes, might the typical magazine publisher with only a freshman course in business mathematics but no econometrics experience inordinate difficulties in determining the signs *and* magnitudes of the slopes *and* intercepts of the demand *and* supply curves?

4. SUPPLY & DEMAND ANALYSIS HAS USES OTHER THAN OPTIMAL PRICING. While the alleged application of optimal pricing to maximizing profit may help in getting the attention and tuition of greedy business majors during one (repeatable) business mathematics course, the underlying supply & demand analy-

sis has other uses, for instance, to public policy:

(A) Health care pricing **[18, 24]**: if the government regulates the price of health care, what consequences may result if it sets price equal to average cost versus marginal cost?
(B) Law enforcement **[14, 15, 25]**: in enforcing laws about illegal gambling, heroin, marijuana, and prostitution, how could the supply and demand affect the government's efficacy?
(C) External evaluation of management **[11, 16]**: do the managers of the Royal Shakespeare Theatre set ticket prices to maximize profit, maximize gross revenues, or maximize attendance? (Non-profit organizations may strengthen their requests for donations and subsidies if they can demonstrate that they maximize not profit but productivity.)
(D) Environment **[10, 15]**: does the population of salmon exhibit a greater sensitivity to habitat protection or to fishery regulation?
(E) Taxes **[6, 17, 20]**: what happens in politicians' minds when they draft new tax laws?

Unfortunately, as for optimal pricing, provably successful examples of applications of supply & demand analysis to public policy seem scarce.

5. CASE STUDY: DO MATHEMATICAL SOCIETIES OPTIMIZE? Though Professor Dudley's analysis indicates that publishers might not set subscription rates to maximize revenues, the same analysis backward lends itself to the solution of an inverse problem, in the manner of Gapinksi **[11]**: with the prevailing rates given, what are the publishers trying to optimize? As an instance of such a reverse analysis, consider Professor Dudley's model,

$$t = s\left(1 + \frac{p}{1+i} - \frac{(c/s)p}{1+i}\right)$$

which equates the present value t of a two-year subscription (paid today) to the present value of two consecutive one-year subscriptions (the first one paid today and the second one paid one year hence), where

c represents the cost of obtaining one new subscription (advertising),
i represents the effective annual rate of interest available for the publisher's investments,
p represents the probability that a one-year subscriber will renew a subscription, and
s represents the price of a one-year subscription.

For one major mathematical society (which will remain anonymous here) and for the same period 1993–1994, an informal search showed that about 2% of the institutional academic members (universities) who subscribed in 1993 did not renew their subscriptions for 1994, which leads to the estimate $p \approx 98\%$. At the same time, the prevailing effective annual rate of interest, for instance, that of the one-year U.S. Treasury bill 29 July 1993–28 July 1994 **[13]**, amounted to about $i \approx 3.60\%$. From $c \geq 0$ and because t is a decreasing function of c, it follows that

$$t = s\left(1 + \frac{0.98}{1 + 0.036} - \frac{0.98(c/s)}{1 + 0.036}\right) \leq s\left(1 + \frac{0.98}{1 + 0.036}\right) \approx 1.95s$$

whence the society can offer at least a 5% discount for the second year's

subscription. With Professor Dudley's estimate of $c/s \approx 0.1$ for loyal readers, $t = s(1 + 0.98[1 - 0.1]/[1 + 0.036]) \approx 1.85s$, which would mean that the society could offer a 15% discount for the second year's subscription. With either discount, if the demand for mathematics magazines slopes downward, in other words, if mathematics magazines are not Giffen goods like 19th century Irish potatoes, then the discounted rates would attract more subscribers. Why then does the society not offer any such discount?

The foregoing backward analysis suggests that the society's management might prefer a higher profit to a larger readership. Also available to outside auditors, as in **[11]**, the same analysis might make a society feel hard-pressed to defend its non-profit status...

6. BUSINESS AND ACADEMIA REMAIN UNAWARE THAT MATHEMATICS CAN HELP OTHERWISE. The preceding considerations indicate that Professor Dudley's problem—the irrelevance of mathematics to price setting—arises from at least two causes: the practical impossibility of verifying the hypotheses (estimates of supply & demand), and the inability of people to apply mathematics (calculations of present values) that their alma mater certify they have the ability to use. While the first cause, lying in econometric difficulties, belongs to the professional realm of economists, the second cause rests solely with the teaching of business mathematics.

That second cause—the inability of people to use the mathematics that we have "taught" them—has been crisply identified by at least one business practitioner, namely CPA Gail A. Eisner **[9]**. Owner of an accounting practice, Eisner describes, among other real situations, the following scene

> [...] a very bright, young CPA [partner] called me into his office and asked how a person could possibly calculate a bonus if the company's formula required that the bonus be 15 percent of the net profit *after* the bonus has been deducted! I showed him the linear equation $B = 0.15(P - B)$ [...] He was flabbergasted. [Italics as in the original.]

For advocates and instructors of undergraduate business mathematics, the revelation ought to lie in the apparent absence of any trace of such business mathematics in an accounting firm's *partner*. Hence arises the question whether business mathematics has a positive net present value to the business community. Eisner's article **[9]** may provide a start for an investigation. Even with a positive answer, however, how to convince business students in the classroom that mathematics will later be able to save them time, hence money, remains an open problem.

Besides elementary algebra, business also has uses for higher mathematics, but again for purposes other than optimal pricing. For example, Northwest Airlines reportedly benefits greatly from the use of linear programming to schedule crews **[3]**, American Airlines reportedly saved \$1.4 *billion* through mathematical programming "over a three-year period in which its net profits were only \$892 million" **[7]**, and investment firms reportedly gain (or loose) money in managing the risks of portfolios through stochastic differential equations **[4, 5]**. Still, such real uses require mathematical and computational sophistication at the graduate level. Hence emerges another relevant question, how to impart just enough mathematics to business majors, so that they may understand the potential power and limitations of mathematics, decide when to hire mathematicians, and consult with them profitably. That, however, also remains an open problem.

Nevertheless, partial solutions exist for the open problems just posed, though at enormous costs in faculty time, money, and talent [23], to incorporate real applications of mathematics in undergraduate courses. Whereas costs so high may still exceed the expected present value of any contract that a publisher might offer to a potential textbook author, schools that are sufficiently serious about the effectiveness—that is, the future value—of the education that they provide may be able to afford one of the least onerous partial solutions: to have mathematicians from industry come and give talks to each class. At the cost of a round-trip ticket and a good hotel room for each speaker, such talks have produced such significant improvements as to induce even underprepared and hitherto unmotivated students to write such evaluations as "I should like to learn enough mathematics to work for that company." Two of the sources of mathematical speakers from industry lie in the MAA's Program of Visiting Lecturers and in SIAM's Visiting Lecturer Program (VLP), organized by Leon H. Seitelman (VLP Chair), Ann K. Stehney (VLP Coordinator), and Gilbert Strang (SIAM Vice President for Education), with information available from vlp@siam.org at SIAM. The bottom line is that if a school does not offer resources so minimal to improve the future value of its mathematics curriculum, then a backward analysis might make it hard-pressed to defend its non-profit status...

REFERENCES

1. American Mathematical Society, *Combined Membership List*, 1994–1995 ed., American Mathematical Society, Providence, RI, 1994.
2. Wayne Angell, "A Single Goal for the Fed," *The Wall Street Journal*, western ed., Vol. CXXXI, No. 97 (Wednesday 16 November 1994), p. A34.
3. Jane Barutt and Theresa Hull, "Airline Crew Scheduling: Supercomputers and Algorithms," *SIAM News*, Vol. 23, No. 6 (November 1990), pp. 1 and 2–22.
4. James Case, "The Mathematization of Finance," *SIAM News*, Vol. 87, No. 6 (July 1994), p. 10.
5. James Case, "The Mathematization of Finance: Part II," *SIAM News*, Vol. 87, No. 7 (August–September 1994), pp. 1 and 12–13.
6. Margaret Cibes, "A Tale of Three Taxes," *The Mathematics Teacher*, Vol. 86, No. 4 (April 1993), pp. 278–281.
7. "Airline Ties Profitability to Yield Management," *SIAM News*, Vol. 27, No. 4 (May–June 1994), pp. 12 and 18–19.
8. Underwood Dudley, "Two-Year Magazine Subscription Rates," *American Mathematical Monthly*, Vol. 100, No. 1 (January 1993), pp. 34–37.
9. Gail A. Eisner, "Using Algebra in an Accounting Practice," *The Mathematics Teacher*, Vol. 87, No. 4 (April 1994), pp. 272–276.
10. Geoffrey T. Evans and J. Brian Dempson, "Calculating the sensitivity of a salmonid population model," *Canadian Journal of Fishery and Aquatic Science*, vol. 43, No. 4 (1986), pp. 863–868.
11. James H. Gapinski, "Do the nonprofit performing arts optimize?" *Quarterly Review of Economics and Business*, Vol. 25, No. 2 (1985), pp. 27–37.
12. Ulrich Kohli, "Robert Giffen and the Irish potato: note," *The American Economic Review*, Vol. 76, No. 3 (1986), pp. 539–542.
13. Thomas D. Lauricella and Constance Mitchell, "Coca-Cola Joins Disney at the Very Long End with the Sale of $150 Million of 100-Year Bonds," *The Wall Street Journal*, western ed., Vol. CXXIX, No. 16 (Friday 23 July 1993), p. C15.
14. Yves Nievergelt, "The Concept of Elasticity in Economics," *SIAM Review*, Vol. 25, No. 2 (April 1983), pp. 261–265.
15. Yves Nievergelt, "Price Elasticity of Demand: Gambling, Heroin, Marijuana, Whiskey, Prositution, and Fish," *U.M.A.P. Journal*, Vol. 8, No. 1 (Spring 1987), pp. 31–61, reprinted as UMAP Module 674, COMAP, Lexington, MA, 1987.
16. Yves Nievergelt, "Executive Mathematics," *U.M.A.P. Journal*, Vol. 8, No. 3 (Autumn 1987), pp. 251–266.
17. Yves Nievergelt, "Functions Give Three Points of View on the New Income Tax Law," *The Mathematics Teacher*, Vol. 81, No. 3 (March 1988), pp. 176–181.

18. Yves Nievergelt, "Graphic Differentiation Clarifies Health Care Pricing," *U.M.A.P. Journal*, Vol. 9, No. 1 (Spring 1988), pp. 51–86, reprinted as UMAP Module 678, COMAP, Lexington, MA, 1988.
19. Yves Nievergelt, "Fisher's Effect: Real Growth is not Interest Less Inflation," *The Mathematics Teacher*, Vol. 81, No. 7 (October 1988), pp. 546–547.
20. Yves Nievergelt, "A Graphic Introduction to Functions: The Federal Income-Tax," *U.M.A.P. Journal*, Vol. 10, No. 2 (Summer 1989), reprinted as UMAP Module 689, COMAP, Lexington, MA, 1989.
21. Yves Nievergelt, *Mathematics in Business Administration*, Richard D. Irwin, Homewood, IL, 1989.
22. Donn R. Pescatrice, "The Inelastic Demand for Wagering," *Applied Economics*, Vol. 12, No. 1 (March 1980), pp. 1–10.
23. Michael C. Reed, "Mainstreaming Calculus Reform," *SIAM News*, Vol. 27, No. 7 (August–September 1994), p. 7.
24. James Studnicki, "Regulation by DRGs: Policy Or Perversion?" *Hospital & Health Services Administration*, Vol. 28, No. 1 (1983), pp. 96–110.
25. Michael E. White and William A. Luksetich, "Heroin: Price Elasticity and Enforcement Strategies," *Economic Inquiry*, Vol. 21, No. 4 (1983), pp. 557–564.

Department of Mathematics, MS-32
Eastern Washington University
526 5th Street
Cheney, WA 99004-2431
ynievergelt@ewu.edu

Mathematics, rightly viewed, possesses not only truth, but supreme beauty—a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the trappings of painting or music, yet sublimely pure, and capable of stern perfection such as only the greatest art can show.

—*Russell*

Answer to Picture Puzzle
(p. 142)

Th.-J. Stieltjes [one of the two only known pictures of him] 1856–1894.

On December 31, 1894, Stieltjes passed away in Toulouse; he was just 38 years old. During the academic year 1994–1995, both in the Netherlands and France events were organized for the commemoration of the 100th anniversary of his death.

Submitted by Professor Jean-Baptiste Hiriart-Urruty
Université Paul Sabatier (Toulouse III)
U.E.R.
Mathématiques, Informatique, Gestion
118, route de Narbonne
F 31062 Toulouse Cédex, FRANCE

Free Elastic Parallels in a Surface of Revolution

Manuel Barros and Oscar J. Garay

1. INTRODUCTION. One of the most classical topics in the Calculus of Variations was proposed by James Bernoulli in 1691: The problem of the bent beam, elastic rod or simply *elastica*. This problem has to do with the following situation: join smoothly the ends of a piano wire and bend the resulting ring so that the wire is held in some configuration described by an immersion of the circle into R^3. Now release the ring and suppose it moves so as to decrease its "bending energy". How does the wire evolve and what will happen ultimately as time goes to infinity? It was Daniel Bernoulli who suggested a model for an elastic rod in equilibrium. Following his model all kinds of elastica should minimize total squared curvature functional, F (also called bending energy functional) among curves of the same length and first order boundary data. In 1743, Euler determined all forms the plane elastic rods may take (see **[T]** for details).

Recently Bryant-Griffiths **[BG]** and Langer-Singer **[LS1, 2]**, have generalized the notion of elastica and studied them from a geometrical point of view. An elastica is a curve in a Riemannian manifold which is a critical point for $F^{\lambda}(\gamma) = \int_{\gamma}(k^2 + \lambda)$, where k is the geodesic curvature of γ, λ is real number and γ is supposed either to be closed or to satisfy given first order boundary data. Geometrically the closed elastica and their global behavior have special interest. In his analysis of the plane elasticae, Euler had already described the closed plane elastica: The circle and the figure eight curve.

In **[LS1]** the term *free elastica* was introduced for the critical points for F among closed curves which are allowed to grow in length (this corresponds to $\lambda = 0$). This lack of constraint of length makes existence an interesting and non trivial question in the Calculus of Variations.

Minimax methods are a useful tool in obtaining critical points for real valued functionals defined on a manifold. However there may not exist any such points and in order to find them one generally needs some kind of "compactness" for the problem. A usual way to guarantee this is the Palais-Smale condition C. Trivial examples of closed free elastica are the closed geodesics. Evidently geodesic circles in a space form are examples of closed non-free elastica. By analyzing explicit solutions for the elastica equation, Langer and Singer classified the closed free elastic curves in 2-dimensional space forms and closed non-free elastic curves in Euclidean spaces **[LS1, 2]**. They also justify, by using a heuristic minimax argument, the existence and instability for all closed non-geodesic free elastica in S^2. In order to transform this heuristic minimax argument into a rigorous one, they need to establish some version of the Palais-Smale condition C. This was made in **[LS3]** for the case of curves in R^3 under the constraint of fixed length. These ideas were generalized in **[LS4]** to prove the existence of closed elastica of fixed length on a compact Riemannian manifold (see also **[K]** for different proofs). The problem

arises when one considers no constraint on arclength. Closed geodesics are "straight" having 0 bending energy. Hence a minimax principle would suggest that between two closed geodesics one could hope to find a minimax closed free elastic. However a Palais-Smale condition is not known to be valid for $\lambda = 0$ **[LS4]**.

In this article we discuss the elasticity of the parallels in a surface of revolution S. It is obvious that any parallel is a closed elastica of S. Thus we will deal with the free elasticity of these curves. In §2 we wonder about surfaces of revolution whose parallels are all free elastica. Then we answer this question by showing that besides right cylinders (all whose parallels are geodesics), the only such surfaces are what we shall call "trumpet" surfaces (which are free of geodesic parallels), (see Figure 1).

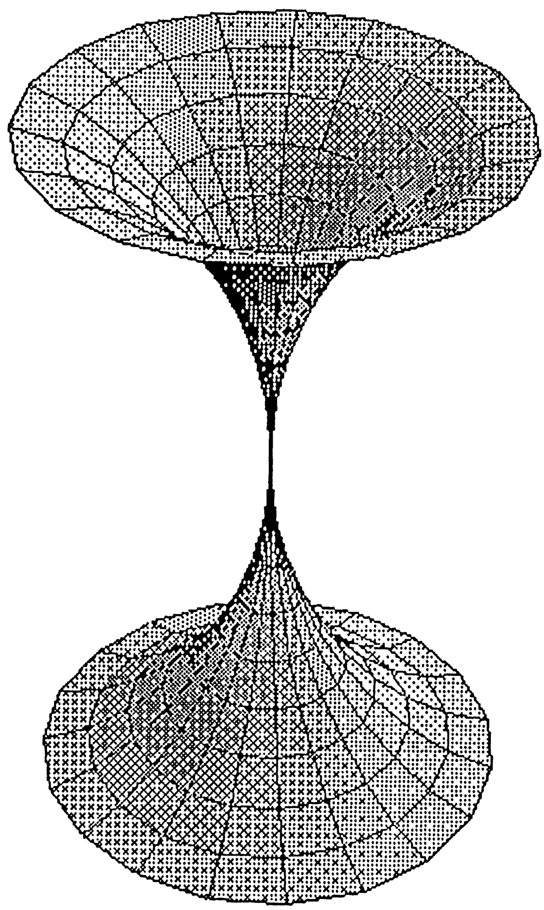


Figure 1

In §3 we study the distribution of the non-geodesic free elastic parallels by using some examples which are meant to show the difficulty of a general analysis. For a given positive function $f(t)$ whose graph is the profile curve of S, one can study $E(t) = [2\pi(f'(t))^2]/[f(t)(1 + (f'(t))^2)]$, the bending energy functional on parallels of S. It is not difficult to see that the critical points of E are the free elastic parallels and the geodesic parallels are the zeroes of E which are global minima. Consequently some kind of one-dimensional variational principle is working here.

Thus a minimax principle would suggest that between two closed geodesics one could hope to find a minimax solution. In our context, Rolle's theorem guarantees the existence of a non-geodesic free elastic parallel between each pair of geodesic parallels. However not every non-geodesic free elastic parallel can be obtained in this way (as Example 1 might indicate). In fact, Example 2 shows this claim.

The first three examples presented share the feature that between each two non-geodesic free elastic parallels, one finds a geodesic parallel (and so a minimum of F). This is not true in general as we show in Example 4. Nevertheless, it is not a counterexample to the following suggested mountain-pass argument but in the other direction: "If non-geodesic free elastic curves are minimax solutions, then perhaps they are separated by a minimum". Just as local maxima of a function of one variable are separated by a minimum. Example 4 occurs because E is strictly positive but has critical points.

Finally, since geodesics are global minima of E, one could expect to find local maxima on either side of them. It is enough to take a geodesic parallel of a non-negative curved surface of revolution and observe that non-geodesic free elastic parallels are made up hyperbolic points, to realize that this is false again (example 5 provides a negatively curved counterexample).

The authors wish to express their gratitude to the referee for his useful comments and suggestions.

2. WHAT IF EVERY PARALLEL IS A FREE ELASTICA? Let $\gamma: [0, L] \to M$ be an immersed curve in an oriented Riemannian surface M. We denote by s and $T(s)$ the arclength and the unit tangent vector of γ respectively. Let $k^2(s) = |\nabla_T T|^2$ be the squared curvature of γ in M, ∇ being the Levi-Civita connection on M. An *elastica* (or *elastic* curve) on M is a curve of M which is critical for the total squared curvature functional $F(\gamma) = \int_\gamma k^2\, ds$ defined on regular curves of a fixed length satisfying given first order boundary data. If we remove the constraint on arclength, we may speak of *free* elastica, (see **[LS1]** for details). By computing the first variation formula for $F(\gamma)$, it was shown in **[LS1]** that the curvature of an elastica must satisfy the following second order differential equation:

$$0 = 2k''(s) + k^3(s) + 2k(s)G(s) - \lambda k(s) \tag{1}$$

where k is the signed curvature of γ, $G(s)$ in the Gaussian curvature of M along γ, k'' denotes the second derivative of k with respect to s and λ is a constant. The case of free elasticae corresponds with $\lambda = 0$.

Remark 1. *It is easily seen from the above equation that a non-geodesic curve γ with constant geodesic curvature k is an elastica, if and only if, the Gaussian curvature of M is constant along γ. In particular, if γ is a free elastica, its points must be hyperbolic points of M.*

From now on we restrict ourselves to the case in which M is a surface of revolution in E^3. Namely M is obtained by rotating a regular plane curve C around an axis in that plane which a priori does not meet the curve. We shall take the *xz-plane* as the plane in which C is contained and the *z-axis* as the rotation axis.

Let $\alpha(s) = (f(s), (g(s))$; $a < s < b$; $f(s) > 0$ be a parameterization by the arclength of C. We denote by u the rotation angle about the *z-axis*. Then each point of C describes a parallel, say γ_s, which can be parameterized by arclength,

say t, in the following way

$$\gamma_s(t) = \left(f(s)\cos\frac{t}{f(s)}, f(s)\sin\frac{t}{f(s)}, g(s)\right) \tag{2}$$

where $0 \leq t \leq 2\pi f(s)$.

First of all we wish to compute the curvature function, say k_s of γ_s in M. The unit tangent vector field to γ_s is

$$T_s(t) = \gamma_s'(t) = \left(-\sin\frac{t}{f(s)}, \cos\frac{t}{f(s)}, 0\right). \tag{3}$$

On the other hand the unit normal vector field, say N, to M along γ_s is given by

$$N(s,t) = \left(-g'(s)\cos\frac{t}{f(s)}, -g'(s)\sin\frac{t}{f(s)}, f'(s)\right). \tag{4}$$

We denote by ∇ the Levi-Civita connection of M and then

$$\nabla_{T_s}T_s(t) = \left(-\frac{(f'(s))^2}{f(s)}\cos\frac{t}{f(s)}, -\frac{(f'(s))^2}{f(s)}\sin\frac{t}{f(s)}, -\frac{f'(s)g'(s)}{f(s)}\right). \tag{5}$$

Now the squared curvature of γ_s into M is

$$(k_s(t))^2 = |\nabla_{T_s}T_s(t)|^2 = \frac{(f'(s))^2}{f^2(s)} \tag{6}$$

which obviously is constant along γ_s.

Remark 2. *Since the Gaussian curvature of S is constant along the parallels, we see from Remark* 1 *that every parallel of a surface of revolution is a closed elastic curve.*

Now we look at the free elastic parallels. We have from equation (1) that γ_s is a free elastica of M if and only if its curvature satisfies the corresponding Euler equation

$$(k_s(t))^3 + 2G(s,t)k_s(t) = 0. \tag{7}$$

Now the Gaussian curvature of M along γ_s is given by (see **[O]** p. 235.)

$$G(s,t) = -\frac{f''(s)}{f(s)}. \tag{8}$$

Therefore we combine (6), (7) and (8) to get that γ_s is a free elastica of M if and only if f satisfies

$$2f(s)f''(s) = (f'(s))^2. \tag{9}$$

Remark 3. *Notice that to obtain* (9) *we have assumed that $f'(s) \neq 0$, otherwise the associated parallel γ_s should be a geodesic of M and so a trivial solution of the Euler equation. Also we deduce from* (7) *that if M has non-negative Gaussian curvature then it does not have non-geodesic free elastic parallels, the converse being non-true, as the bugle surface shows* (*see Remark* 6).

Theorem 1. *Let $M \subset E^3$ be a surface of revolution obtained by rotating the arclength parametrized curve $\alpha(s) = (f(s), g(s))$, with $s \in (-a, a)$, in the xz-plane around the z-axis. Then M has all parallels being free elastica in M if and only if either*:

(1) *f is some positive constant and so M is a right circular cylinder, or*
(2) *there exists a positive number c and*

$$\alpha(s) = \left(\frac{c}{4}s^2, \frac{s}{2}\sqrt{1 - \frac{c^2}{4}s^2} - \frac{1}{c}\arccos\frac{c}{2}s + b\right) \tag{10}$$

with $s \in (-2/c, 0) \cup (0, 2/c)$ and $b \in R$.

Proof: Let us assume M is not a right circular cylinder, then it has all parallels being free elastica iff f is a solution of the following second order differential equation

$$f''(s) = \frac{(f'(s))^2}{2f(s)} \tag{11}$$

and certainly we can use standard arguments to get the integration of the last equation. We begin by doing a first integration to obtain

$$(f'(s))^2 = cf(s) \tag{12}$$

where c is any positive real number. Then the general solution of (11) is

$$f(s) = \frac{c}{4}s^2 \tag{13}$$

where we have normalized by choosing a convenient arclength parameter in order to avoid a nonessential parameter. Now we can compute g from $g'(s) = \sqrt{1 - (f'(s))^2}$ and so we certainly obtain the second coordinate function of $\alpha(s)$ announced in (10) (see Figure 1).

3. SOME EXAMPLES. In order to get nice examples we will consider the curve $\alpha(t) = (h(t), g(t))$, $a < t < b$ and $h(t) > 0$ anywhere, with an arbitrary parameterization. If $g(t)$ is not a constant, then locally we may represent α by $\alpha(t) = (f(t), t)$ where $a < t < b$ and $f(t) > 0$ anywhere. We rotate this curve in the *xz-plane* around the *z-axis* to obtain a surface of revolution M which can be parameterized by

$$X(t, u) = (f(t)\cos u, f(t)\sin u, t). \tag{14}$$

The Gaussian curvature of this surface is given by [**O**, p. 235]

$$G(t, u) = -\frac{f''(t)}{f(t)\left(1 + (f'(t))^2\right)^2}. \tag{15}$$

Now for any $t \in (a, b)$ we consider the parallel

$$\gamma_t(u) = (f(t)\cos u, f(t)\sin u, t). \tag{16}$$

An easy computation shows that the curvature k_t of γ_t in M satisfies

$$(k_t(u))^2 = \frac{(f'(t))^2}{f^2(t)\left(1 + (f'(t))^2\right)}. \tag{17}$$

Consequently γ_t is a free elastica in M if and only if $f(t)$ satisfies

$$2f(t)f''(t) = (f'(t))^2(1 + (f'(t))^2). \tag{18}$$

Remark 4. *According to* (14) *and* (17), *the points of a parallel of M which is a candidate to be a free elastica in M are hyperbolic.*

3.1 Example 1. *A torus of revolution has exactly two non-geodesic parallels being free elastica.* Let us consider the torus of revolution obtained by rotating the circle centered at $(a,0,0)$ and with radius r, $(a > r)$ around the *z-axis*. The region of this torus where the Gaussian curvature is negative, is obtained by rotating the curve $\alpha(t) = (a - \sqrt{r^2 - t^2}, t)$ where $t \in (-r, r)$, in the *xz-plane* around the *z-axis*. Take $\delta = 2a\sqrt{a^2 - r^2} - 2(a^2 - r^2)$, then $0 < \delta < r^2$. It is clear that $t \in \{\sqrt{\delta}, -\sqrt{\delta}\}$ give two solutions of (18) and so two parallels being free elastica. Notice that $t = 0$ gives a geodesic. Also this surface has another parallel being a geodesic but it is made up elliptic points.

3.2 Example 2. *The catenoid has exactly two non-geodesic parallels being free elastica.* The catenoid is the surface of revolution M in E^3 obtained by rotating the curve $\alpha(t) = (a\cosh(t/a + b), t)$ in the *xz-plane* around the *z-axis*. It has a parallel being a geodesic in M, namely the parallel associated with $t = -ab$. Notice that $\alpha(-ab) = (a, a(\cosh^{-1} 1 - b))$. Now we use (18) to see that the parallels associated with

$$\alpha\big(a(\cosh^{-1}\sqrt{3} - b)\big) = \big(a\sqrt{3}, a(\cosh^{-1}\sqrt{3} - b)\big)$$

and

$$\alpha\big(a(-\cosh^{-1}\sqrt{3} - b)\big) = \big(a\sqrt{3}, a(-\cosh^{-1}\sqrt{3} - b)\big)$$

give two non-geodesic free elastic curves in M.

3.3 Example 3. Both examples given above have exactly two non-geodesic free elastic parallels. In order to understand better the distribution of this kind of curves in a surface of revolution, let us exhibit an example of a surface of revolution which has an arbitrary number of free elastic parallels.

Let us consider the curve $\alpha(t) = (f(t), t)$ in the *xz-plane* where $f(t) = 2 + \sin t$ and $t \in R$. By rotating this around the *z-axis*, one obtains a surface of revolution whose free elastic parallels, according to (18), correspond to the values of t which satisfy the equation

$$H(t) = 0 \tag{19}$$

where $H(t) = \sin^4 t - \sin^2 t + 4\sin t + 2$. Since $\lim_{t\to\pi} H(t) > 0$ and $\lim_{t\to 3\pi/2} H(t) < 0$ one gets that (19) has a solution in $I_1 = (\pi, 3\pi/2)$ and so a free elastic parallel. Also we notice that $H'(t) < 0$ in I_1 and so we obtain exactly one free elastic parallel in I_1. The same argument can be used to find exactly one free elastic parallel in $I_2 = (3\pi/2, 2\pi)$. Now we move I_1 and I_2 an integer multiple of 2π in R to get an arbitrary number of free elastic parallels, (see Figure 2).

Remark 5. *In the given examples we notice that between each two free elastic parallels one finds a geodesic parallel. One might think that this is true in general, however we shall see that this is not the case. In order to demonstrate this claim, we will give an*

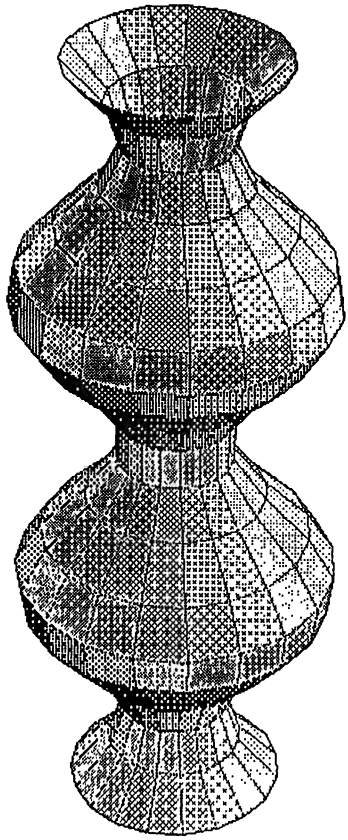


Figure 2

example of a surface of revolution which is free of geodesic parallels and it has an arbitrary number of non-geodesic free elastic parallels.

3.4 Example 4. Let us consider the curve $\alpha(t) = (f(t), t)$ in the *xz-plane* where $f(t) = 2t - \cos t$ and $t > \frac{1}{2}$ (see Figure 3). By rotating this around the *z-axis* we get a surface of revolution, say M, which has not any parallel being a geodesic because $f'(t) = 2 + \sin t > 0$. However and according to (18), the free elastic

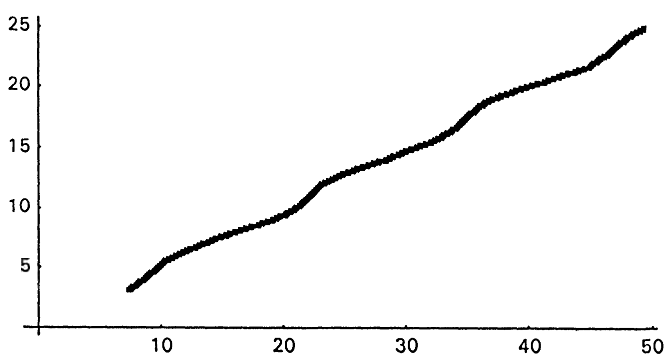


Figure 3

parallels of M correspond to the solutions of

$$F(t) = 0 \tag{20}$$

where $F(t) = \sin^4 t + 8\sin^3 t + 23\sin^2 t + 36\sin t + 22 - 4t\cos t$. Now $F(3\pi/2) > 0$ and $F(2\pi) < 0$ which proves that (20) has a solution in $(3\pi/2, 2\pi)$. Also (20) has another solution in $(2\pi, 5\pi/2)$. Then we move both intervals a natural multiple of 2π in $t > \frac{1}{2}$ to get an arbitrary number of non-geodesic free elastic parallels.

3.5 Example 5. We consider the surface of revolution obtained by rotating around the *z-axis* the curve $\alpha(s) = (f(s), g(s))$ in the *xz-plane* given by $f(s) = 1 + s^2$, $g(s) = \int\sqrt{1 - 4s^2}\, ds$, where s is arclength parameter running in $(-\frac{1}{2}, \frac{1}{2})$. It is not difficult to see that we have a negative curved surface which has a geodesic parallel associated to $s = 0$ but it does not have any non-geodesic free elastic parallel.

Remark 6. *The hyperbolic plane is very rich in free elastic curves, even closed free elastic curves,* **[LS1]**. *One way to see locally this surface in E^3 as a surface of revolution is by means of the* bugle *surface (also called the tractroid or the pseudosphere). However this surface has not free elastic parallels.*

REFERENCES

[BG] R. Bryant and P. Griffiths. Reduction for constrained variational problems and $\int(k^2/s)\,ds$. *Amer. J. Math.* 108 (1986), 525–570.

[K] N. Koiso. *Convergence towards an elastica.* Geometry and Global Analysis. Report of the first MSJ International Research Institute, 1993, Tohoku Univ., 249–255.

[LS1] J. Langer and D. A. Singer. The total squared curvature of closed curves. *Journal Diff. Geom.* 20 (1984), 1–22.

[LS2] J. Langer and D. A. Singer. Knotted elastic curves in R^3. *Journal London Math. Soc.* 30 (1984), 512–520.

[LS3] J. Langer and D. A. Singer. Curve straightening and a minimax argument for closed elastic curves. *Topology* 24 (1985), 75–88.

[LS4] J. Langer and D. A. Singer. Curve-straightening in Riemannian manifolds. *Annals of Global Anal. and Geom.* 5 (1987), 133–150.

[O] B. O'Neill. *Elementary Differential Geometry*, Academic Press (1966).

[T] C. Truesdell. The influence of elasticity in Analysis. The classical heritage. *Bull. A.M.S.* 9, (1983), 293–310.

Departamento de Geometria y Topologia
Facultad de Ciencias
Universidad de Granada
18071 Granada, SPAIN
mbarros@goliat.ugr.es

Departamento de Matematicas
Universidad del Pais Vasco
Aptdo 644
48080 Bilbao, SPAIN
mtpgabeo@lg.ehu.es

NOTES

Edited by: **John Duncan**

A Simple Proof of the Jordan Decomposition Theorem for Matrices

Israel Gohberg and Seymour Goldberg

There are several proofs of the existence of the Jordan normal form of a square complex matrix. For a discussion of these proofs and references, we refer the reader to the paper by Väliaho [**1**].

In this note we present a new, simple and short proof of the existence of the Jordan decomposition of an operator on a finite dimensional vector space over the complex numbers. Our proof is based on an algorithm that allows one to build the Jordan form of an operator A on an n-dimensional space if the Jordan form of A restricted to an $n-1$ dimensional invariant subspace is known.

Let A be a linear operator on a finite-dimensional vector space V over the complex numbers. Recall that a subspace of V is called *cyclic* if it is of the form

$$\operatorname{span}\{\varphi, (A-\lambda)\varphi, \dots, (A-\lambda)^{m-1}\varphi\}$$

with $(A-\lambda)^{m-1}\varphi \neq 0$ and $(A-\lambda)^m\varphi = 0$. Such a subspace is A-invariant and has dimension m. This follows immediately from the fact that if for some $r (r = 0, 1, \dots, m-1)$

$$c_r(A-\lambda)^r\varphi + \cdots + c_{m-1}(A-\lambda)^{m-1}\varphi = 0 \quad \text{and} \quad c_r \neq 0,$$

then after an application of $(A-\lambda)^{m-r-1}$ to both sides of this equality we obtain

$$c_r(A-\lambda)^{m-1}\varphi = 0.$$

Idea of the proof: The argument can be reduced to two cases. In one case there is a vector g outside of an $n-1$ dimensional A-invariant subspace F of V such that $Ag = 0$. In this case $V = F \oplus \operatorname{span}\{g\}$ and the solution is clear from the induction hypothesis on F. The difficult case is when no such g exists. It turns out that one of the cyclic subspaces of the restriction of A to F is replaced by a cyclic subspace of A in V which is larger by one dimension while keeping the other cyclic subspaces unchanged.

Observation. Suppose $W = H \oplus \operatorname{span}\{\varphi, A\varphi, \dots, A^{m-1}\varphi\}$ with $A^{m-1}\varphi \neq 0$, $A^m\varphi = 0$, where H is an A-invariant subspace of V and $A^mH = \{0\}$. Given $h \in H$, let $\varphi' = \varphi + h$. Then

$$W = H \oplus \operatorname{span}\{\varphi', A\varphi', \dots, A^{m-1}\varphi'\},$$

with $A^{m-1}\varphi' \neq 0$ and $A^m\varphi' = 0$. This statement follows immediately from the fact that if a linear combination of the vectors $\varphi', A\varphi', \dots, A^{m-1}\varphi'$ belongs to H, then the same linear combination of the vectors $\varphi, A\varphi, \dots, A^{m-1}\varphi$ also belongs to H.

Jordan Decomposition Theorem. *Let $V \neq (0)$ be a finite dimensional vector space over the complex numbers and let A be a linear operator on V. Then V can be expressed as a direct sum of cyclic subspaces.*

Proof: The proof proceeds by induction on $\dim V$. The decomposition is trivial if $\dim V = 1$. Assume that the decomposition holds for spaces of dimension $n - 1$. Let $\dim V = n$. First we assume that A is singular. Then the range $R(A)$ of A has dimension at most $n - 1$. Let F be an $n - 1$ dimensional subspace of V which contains $R(A)$. Since $AF \subset R(A) \subset F$, the induction hypothesis guarantees that F is the direct sum of cyclic subspaces

$$M_j = \operatorname{span}\left\{\varphi_j, (A - \lambda_j)\varphi_j, \dots, (A - \lambda_j)^{m_j - 1}\varphi_j\right\}, \quad 1 \leq j \leq k.$$

The subscripts are chosen so that $\dim M_j \leq \dim M_{j+1}$, $1 \leq j \leq k - 1$. Define $S = \{j \mid \lambda_j = 0\}$. Take $g \notin F$. We claim that Ag is of the form

$$Ag = \sum_{j \in S} \alpha_j \varphi_j + Ah, \quad h \in F, \tag{1}$$

if $S \neq \varnothing$. If $S = \varnothing$, then $Ag = Ah$. To verify (1), note that $Ag \in R(A) \subset F$. Hence Ag is a linear combination of vectors of the form $(A - \lambda_j)^q \varphi_j$, $0 \leq q \leq m_j - 1$, $1 \leq j \leq k$. For $\lambda_j = 0$, the vectors $A\varphi_j, \dots, A_{\varphi_j}^{m-1}$, are in $A(F)$. If $\lambda_j \neq 0$, then from $(A - \lambda_j)^{m_j}\varphi_j = 0$ and the binomial decomposition we get that φ_j is of the form $\sum_{m=1}^{m_j} b_m A^m \varphi_j$. Thus all vectors $(A - \lambda_j)^q \varphi_j$ belong to $A(F)$ and equation (1) holds.

Let $g_1 = g - h$, where h is given in (1). Since $g \notin F$ and $h \in F$, $g_1 \notin F$ and from equation (1),

$$Ag_1 = \sum_{j \in S} \alpha_j \varphi_j. \tag{2}$$

If $Ag_1 = 0$, then $\operatorname{span}\{g_1\}$ is cyclic and $V = F \oplus \operatorname{span}\{g_1\}$. Suppose $Ag_1 \neq 0$. Let p be the largest of the integers j in (2) for which $\alpha_j \neq 0$. Then for $\tilde{g} = (1/\alpha_p)g_1$,

$$A\tilde{g} = \varphi_p + \sum_{j \in S,\, j < p} \frac{\alpha_j}{\alpha_p} \varphi_j. \tag{3}$$

Define

$$H = \sum_{j \in S,\, j < p} \oplus M_j.$$

The subspace H is A-invariant and since $\dim M_j \leq \dim M_p$, $j < p$, it follows that $A^{m_p}(H) = \{0\}$. Thus by the observation applied to $H \oplus M_p$ and equality (3), we have

$$H \oplus M_p = H \oplus \operatorname{span}\{A\tilde{g}, \dots, A^{m_p}\tilde{g}\}.$$

Hence,

$$F = \sum_{j \neq p} \oplus M_j \oplus \operatorname{span}\{A\tilde{g}, \dots, A^{m_p}\tilde{g}\}.$$

Since $\tilde{g} \notin F$,

$$V = F \oplus \operatorname{span}\{\tilde{g}\} = \sum_{j \neq p} \oplus M_j \oplus \operatorname{span}\{\tilde{g}, A\tilde{g}, \dots, A^{m_p}\tilde{g}\}.$$

This completes the proof of the theorem under the assumption that A is singular. For the general case, let μ be an eigenvalue of A. Then $A - \mu$ is

singular and by the above result applied to $A - \mu$, it follows that V is the direct sum of cyclic subspaces for A. Q.E.D.

This proof shows how to extend a Jordan form for A on an $n - 1$ dimensional invariant subspace F to an n-dimensional A-invariant subspace containing F.

Note that the proof of the theorem also holds if the scalars of complex numbers is replaced by an algebraically closed field.

Illustrative Example. Let

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & a \\ 0 & 0 & 0 & 0 & b \\ 0 & 0 & 0 & 1 & c \\ 0 & 0 & 0 & 0 & d \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$Ae_2 = e_1, \quad Ae_1 = 0, \quad Ae_4 = e_3, \quad Ae_3 = 0.$$

We take

$$F = \operatorname{span}\{e_1, e_2, e_3, e_4\} = \operatorname{span}\{e_2, Ae_2\} \oplus \operatorname{span}\{e_4, Ae_4\}.$$

Now $e_5 \notin F$ and

$$Ae_5 = ae_1 + be_2 + ce_3 + de_4 = be_2 + de_4 + A(ae_2 + ce_4).$$

If $d \neq 0$, take $\tilde{g} = e_5 - ae_2 - ce_4/d$. Then $A\tilde{g} = e_4 + (b/d)e_2$, $A^2\tilde{g} = e_3 + (b/d)e_1$ and

$$\mathbb{C}^5 = \operatorname{span}\{e_2, Ae_2\} \oplus \operatorname{span}\{\tilde{g}, A\tilde{g}, A^2\tilde{g}\}.$$

If $d = 0$ and $b \neq 0$ take $\tilde{g} = e_5 - ae_2 - ce_4/b$. Then $A\tilde{g} = e_2$ and $Ae_2 = e_1$. Hence

$$\mathbb{C}^5 = \operatorname{span}\{\tilde{g}, A\tilde{g}, A^2\tilde{g}\} \oplus \operatorname{span}\{e_4, Ae_4\}.$$

Finally, if $d = b = 0$ take $\tilde{g} = e_5 - ae_2 - ce_4$. Then $A\tilde{g} = 0$ and

$$\mathbb{C}^5 = \operatorname{span}\{e_2, Ae_2\} \oplus \operatorname{span}\{e_4, Ae_4\} \oplus \operatorname{span}\{\tilde{g}\}.$$

REFERENCE

1. H. Väliaho, An elementary approach to the Jordan form of a matrix, this Monthly, 93, No. 9, (1986) 711–714.

Raymond and Beverly Sackler Faculty of Exact Sciences
School of Mathematical Sciences
Tel Aviv University
Tel Aviv 69978
Israel

Department of Mathematics
University of Maryland
College Park, MD 20742

A Concise Proof of Hilbert's Basis Theorem

A. Caruth

Let R denote a commutative Noetherian ring with an identity element. Hilbert's basis theorem states that the polynomial ring in a finite number of indeterminates over R is also Noetherian. The original theorem in **[3]** is stated for the case when R is a field or the ring of integers. The standard proofs of this fundamental theorem, both for the commutative and non-commutative cases, are essentially of a direct type e.g. (Zariski and Samuel **[6]**, theorem 1, p. 201) where two proofs are given, one of which is substantially that given by Hilbert: see also **[1, 4, 5]**. The present note offers an indirect (i.e. reductio ad absurdum) proof using the fact that in a non-Noetherian ring the set of ideals which are not finitely generated forms, under the relation of inclusion, a partially ordered inductive set: by Zorn's Lemma this set has a maximal element. Since $R[X_1, X_2, \dots, X_n] = R[X_1, X_2, \dots, X_{n-1}][X_n]$ it is sufficient, by induction, to consider the case of a polynomial ring S in a single indeterminate X over R.

When B is an ideal of S, recall that $B : SX = \{c \in S \mid cX \in B\}$ is an ideal of S containing B.

Hilbert's Basis Theorem. *If R is Noetherian, then $S = R[X]$ is also Noetherian.*

Proof: Assume that S is not Noetherian and let B denote an ideal maximal among ideals of S which are not finitely generated. Since $R \cong S/SX$ it follows that $X \notin B$, otherwise B/XS (and therefore B) is finitely generated. Thus the ideal $B + SX$ properly contains B and is therefore finitely generated. Hence, a finitely generated ideal $A = Sa_1 + Sa_2 + \cdots + Sa_n \subset B$ exists such that $B + SX = A + SX$ and so

$$B = A + B \cap SX = A + (B : SX)X.$$

If $B : SX$ properly contains B, then $B : SX$ (and therefore $(B : SX)X$) is finitely generated implying that B is finitely generated.

Hence, $B : SX = B$ and we obtain

$$B = A + BX = A + BX^2 = \cdots = A + BX^s \text{ for every } s \geq 1. \tag{1}$$

Let $k = \max\{\deg a_1, \deg a_2, \dots, \deg a_n\}$. The elements of B of degree $\leq k - 1$ form a submodule C of the R-module generated by $1, X, X^2, \dots, X^{k-1}$ and so C is finitely generated over R (**[6]**, theorem 18, p. 158). Let $f \in B$ with $\deg f = m$ (say). Using equation (1), f can be expressed in the form $\sum_{i=1}^{n} s_i a_i + hX^{m+1}$ where $s_i \in S$ with $\deg s_i \leq m$ and $h \in B$. Thus, $\deg h \leq m + k - (m + 1) = k - 1$ on comparing coefficients of X^j. Accordingly, $h \in C$ and $f \in A + CX^{m+1} \subseteq A + CS \subseteq B$. We now have the contradiction that $B = A + CS$ is finitely generated. Hence, S is Noetherian.

REFERENCES

1. P. B. Bhattacharya, S. K. Jain and S. R. Nagpaul, *Basic Abstract Algebra*, Cambridge University Press, 1986.
2. A. Caruth, Unified proofs of Hilbert's basis theorem and its analogue in formal power series rings, *Math. Proc. Cambridge Phil. Soc.*, 101 (1987), p. 207.
3. D. Hilbert, Ueber die theorie der algebraischen formen, *Math. Ann.* 36 (1890), pp. 473–534.
4. N. Jacobson, *Basic Algebra II*, 2nd ed., W. H. Freeman and Co., 1989.
5. S. Lang, Algebra, 3rd ed., Addison-Wesley Pub. Co., 1993.
6. O. Zariski and P. Samuel, *Commutative algebra*, vol. 1, D. Van Nostrand, Princeton, 1958.

School of Mathematical Studies
University of Portsmouth
Portsmouth, PO1 2EG
England

Rolle's Theorem Fails in l_2

Jesús Ferrer

In [1, Theorems 1 and 2] two multidimensional versions of Rolle's theorem are given which have the classical one-dimensional result as a particular case. Simplifying their statements and notation, let B, U and S denote the closed unit ball, open unit ball and unit sphere, respectively, of $\mathbb{R}^n$. We reproduce the results before mentioned.

Theorem 1. *Let $f: B \to \mathbb{R}^p$ be a continuous function differentiable in U. Assume there is a vector $v \in \mathbb{R}^p$ such that $\langle v, f(x)\rangle = 0$, for every $x \in S$. Then there is a vector $x_0 \in U$ such that $\langle v, f'(x_0)u\rangle = 0$, for all $u \in \mathbb{R}^n$.*

Theorem 2. *Let f be as before. Let $v \in \mathbb{R}^p$ and $z \in U$ be such that $\langle v, f(x) - f(z)\rangle$ does not change sign in S. Then there is a vector $x_0 \in U$ such that $\langle v, f'(x_0)u\rangle = 0$, for all $u \in \mathbb{R}^n$.*

The paper ends with the conjecture that both theorems should not hold for infinite-dimensional domains. We prove the conjecture to be correct by means of an example of a real valued function f defined in the Hilbert space l_2 of square-summable real sequences such that it is continuous and differentiable in every point of l_2, $f_{|S} = 0$ but $f'(x) \neq 0$ for every $x \in U$. Clearly, from now on B, U and S will refer to the closed unit ball, open unit ball and unit sphere, respectively, of l_2. We use $\langle \cdot, \cdot \rangle$ to denote the usual inner product of l_2.

The Example. Let L and R denote the continuous linear operators in l_2 given by, if $x = (x_1, x_2, x_3, \dots)$,

$$Lx = (x_2, x_3, x_4, \dots),$$
$$Rx = (0, x_1, x_2, x_3, \dots).$$

Let T be the map (clearly motivated by [1, Example 1]) $T: l_2 \to l_2$ defined as

$$T(x) = \left(1/2 - \|x\|^2\right)e_1 + Rx.$$

Finally, consider the function $f\colon l_2 \to \mathbb{R}$ such that

$$f(x) = \frac{1 - \|x\|^2}{\|x - T(x)\|^2}.$$

Since the map T has no fixed points, it follows that f is continuous in l_2 and $f(x) = 0$ for every $x \in S$. We show next that f is differentiable in every point of l_2.

Identifying in the usual fashion l_2 with its dual, we know that the Fréchet derivative of $\|x\|^2$ is given by $2x$. So, we have that the mapping T is differentiable at x and, for each $u \in l_2$,

$$T'(x)u = -2\langle x, u\rangle e_1 + Ru.$$

Hence, we have that the derivative of $\|x - T(x)\|^2$ is given by the functional

$$u \to 2\langle x - T(x), u - T'(x)u\rangle.$$

Now since $\|x - T(x)\|^2$ never vanishes, the derivative of a quotient tells us that f is Fréchet differentiable at every $x \in l_2$ and, for each $u \in l_2$, we have

$$f'(x)u = \frac{1}{\|x - T(x)\|^4} \times \left[-2\|x - T(x)\|^2\langle x, u\rangle - 2(1 - \|x\|^2)\langle x - T(x), u - T'(x)u\rangle\right].$$

But, since $\langle T(x), e_1\rangle = 1/2 - \|x\|^2$ and noticing that $\langle x, Ru\rangle = \langle Lx, u\rangle$, $LT(x) = x$, it follows that

$$\begin{aligned}
&\langle x - T(x), u - T'(x)u\rangle \\
&\quad= \langle x - T(x), u\rangle + 2\langle x, u\rangle x_1 - 2\langle x, u\rangle(1/2 - \|x\|^2) - \langle x - T(x), Ru\rangle \\
&\quad= \left\langle x - T(x) + 2x_1x - (1 - 2\|x\|^2)x - L(x - T(x)), u\right\rangle \\
&\quad= \left\langle\left(1 + 2x_1 + 2\|x\|^2\right)x - T(x) - Lx, u\right\rangle.
\end{aligned}$$

Therefore, the value of $f'(x)u$ is given by the expression

$$\frac{-2}{\|x - T(x)\|^4} \times\left\langle\left(\|x - T(x)\|^2 + (1 - \|x\|^2)(1 + 2x_1 + 2\|x\|^2)\right)x - (1 - \|x\|^2)(Lx + T(x)), u\right\rangle.$$

That is,

$$f'(x) = \frac{-2}{\|x - T(x)\|^4} \times\left[\left(\|x - T(x)\|^2 + (1 - \|x\|^2)(1 + 2x_1 + 2\|x\|^2)\right)x - (1 - \|x\|^2)(Lx + T(x))\right].$$

We show that the equation $f'(x) = 0$ has no solution in U. Assume that $f'(x) = 0$, $\|x\| < 1$. Then, if we call

$$s = \frac{\|x - T(x)\|^2}{1 - \|x\|^2} + 1 + 2x_1 + 2\|x\|^2, \tag{1}$$

it follows that

$$Lx + T(x) = sx,$$

and

$$L^2x - sLx + x = 0.$$

That is, $x \in \operatorname{Ker}(L^2 - sL + I)$ is a recurrent sequence of order two in l_2. The associated characteristic equation for this type of sequence is

$$t^2 - st + 1 = 0,$$

which gives us three different alternatives according to the sign of its discriminant.

Case 1. $|s| = 2$. Then we know that the sequences

$$u = \left(1, s/2, (s/2)^2, (s/2)^3, \dots\right); \quad v = \left(0, s/2, 2(s/2)^2, 3(s/2)^3, \dots\right)$$

are basic elements of $\operatorname{Ker}(L^2 - sL + I)$. Thus, $x = Au + Bv$, for some real numbers A, B. So, for each $n \geq 1$,

$$x_n = A(s/2)^{n-1} + B(n-1)(s/2)^{n-1},$$

and, since $\lim_n x_n = 0$, we have that $A = B = 0$, i.e., $x = 0$. But this cannot be so, since

$$f'(0) = 16e_1.$$

Case 2. $|s| < 2$. Then the characteristic equation has two complex roots given by

$$\alpha = \cos\theta + i\sin\theta, \quad \beta = \cos\theta - i\sin\theta, \quad \sin\theta \neq 0.$$

Then, we know that there are complex constants A, B for which

$$x_n = A(\cos\theta + i\sin\theta)^{n-1} + B(\cos\theta + i\sin\theta)^{n-1}, \quad n \geq 1,$$

and, for suitable real constants C, D

$$x_n = C\cos(n-1)\theta + D\sin(n-1)\theta, \quad n \geq 1.$$

But $\sin\theta \neq 0$ implies that the former sequence has no limit, unless $C = D = 0$, i.e., $x = 0$, again a contradiction.

Case 3. $|s| > 2$. We then have two real roots

$$\alpha = \frac{s + \sqrt{s^2 - 4}}{2}, \quad \beta = \frac{s - \sqrt{s^2 - 4}}{2}.$$

Clearly, one of these roots has absolute value greater than one and the other less than one. Assume that

$$|\alpha| > 1, \quad |\beta| < 1.$$

Since

$$x_n = A\alpha^{n-1} + B\beta^{n-1}, \quad n \geq 1,$$

it follows that $A = 0$ and

$$x_n = x_1\beta^{n-1}, \quad n \geq 1.$$

Thus, x is the geometric progression

$$(x_1, x_1\beta, x_1\beta^2, x_1\beta^3, \dots),$$

$$\|x\|^2 = \frac{x_1^2}{1-\beta^2}, \quad \|x - T(x)\|^2 = \left(x_1 + \frac{x_1^2}{1-\beta^2} - \frac{1}{2}\right)^2 + \frac{x_1^2(1-\beta)}{1+\beta}.$$

From $sx = T(x) + Lx$, we have

$$x_1^2 + \frac{1-\beta^2}{\beta}x_1 - \frac{1}{2}(1-\beta^2) = 0, \tag{2}$$

and

$$\|x - T(x)\|^2 = \frac{x_1^2(1-\beta)}{\beta^2(1+\beta)}.$$

Hence, substituting in (1),

$$\beta + \frac{1}{\beta} = s = \frac{x_1^2(1-\beta)}{\beta^2(1+\beta)} \cdot \frac{1-\beta^2}{1-\beta^2-x_1^2} + 1 + 2x_1 + 2\frac{x_1^2}{1-\beta^2},$$

which yields

$$1 = (\beta - 2x_1)\left(1 + \frac{(1-\beta^2)(1-\beta)}{2\beta^2(1-x_1^2-\beta^2)}\right). \tag{3}$$

From (2), we consider two subcases:

(3.1) $$x_1 = \frac{-1+\beta^2-\sqrt{1-\beta^4}}{2\beta}$$

From (3), since $\|x\| < 1$ implies $x_1^2 + \beta^2 < 1$, we have that $0 < \beta - 2x_1 < 1$. Therefore, $0 < (1 + \sqrt{1-\beta^4})/\beta < 1$. A contradiction, since $|\beta| < 1$.

(3.2) $$x_1 = \frac{-1+\beta^2+\sqrt{1-\beta^4}}{2\beta}$$

Noticing that

$$1 - x_1^2 - \beta^2 = \frac{1}{2\beta}(1-\beta^2)(\beta + 2x_1),$$

it follows after (3) that

$$1 = \frac{1}{\beta}\left(1 - \sqrt{1-\beta^4}\right)\frac{2\beta^2 - \beta + \sqrt{1-\beta^4}}{2\beta^2 - 1 + \sqrt{1-\beta^4}},$$

$$\beta\left(2\beta^2 - 1 + \sqrt{1-\beta^4}\right) = \left(1 - \sqrt{1-\beta^4}\right)\left(2\beta^2 - \beta + \sqrt{1-\beta^4}\right)$$

$$2\beta^3 = \left(1 - \sqrt{1-\beta^4}\right)\left(2\beta^2 + \sqrt{1-\beta^4}\right)$$

$$2\left(1 + \sqrt{1-\beta^4}\right) = \beta\left(2\beta^2 + \sqrt{1-\beta^4}\right)$$

$$2(1-\beta^3) = (\beta - 2)\sqrt{1-\beta^4}.$$

This last expression is a contradiction.

REFERENCE

1. M. Furi, and M. Martelli, A Multidimensional Version of Rolle's Theorem. *Am. Math. Monthly*, 102 (1995), 243–249.

Departamento de Anâlisis Matemâtico
Universidad de Valencia
Dr. Moliner, 50
46100 Burjasot (Valencia)
Spain

We have heard much about the poetry of mathematics, but very little of it has as yet been sung. The ancients had a juster notion of their poetic value than we. The most distinct and beautiful statements of any truth must take at last the mathematical form. We might so simplify the rules of moral philosophy, as well as of arithmetic, that one formula would express them both.

—*H. D. Thoreau*

UNSOLVED PROBLEMS

Edited by: **Richard Guy & Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Self-Reading Sequences

Gheorghe Păun and Arto Salomaa

We denote by V^* the set of all words and by V^+ the set of all nonempty words over the alphabet V. The length of a word x is denoted by $|x|$ and $|x|_a$ is the number of occurrences in x of $a \in V$. The set of infinite sequences over V is denoted by V^ω.

Starting from a word $x \in V^+$ and a mapping $f: V^+ \to V^+$ we can construct the sequence

$$z(f,x) = x_0 x_1 x_2 \dots,$$

where $x_0 = x$ and $x_{n+1} = f(x_0 x_1 \dots x_n)$, $n \geq 1$.

Such sequences were first considered in [**11**], for $V = \{0,1\}$ and f one of the following mappings (in all cases $x \in V^+$; $\beta_2(n)$ is the representation of n in base 2):

$$f_{2,0}(x) = \beta_2(|x|_0)0, \qquad f_{2,1}(x) = \beta_2(|x|_1)1,$$
$$f'_{2,0}(x) = \beta_2(|x|_0), \qquad f'_{2,1}(x) = \beta_2(|x|_1).$$

Examples (the parentheses are added for clarity only):

$$z(f_{2,0},1) = 1(0)0(10)0(100)0(111)0(1000)0(1100)0(1111)0\dots$$
$$z(f_{2,1},1) = 1(1)1(11)1(110)1(1001)1(1100)1(1111)1(10100)1\dots$$
$$z(f'_{2,0},100) = 100(10)(11)(11)(11)\dots$$

These sequences were called *self-reading* following the suggestion (and the motivation) in [**3**], [**4**]. Please note that the mappings $f_{2,i}, f'_{2,i}$ above "read" the whole word (counting the number of occurrences of i) and express the output in base 2 and attach one more i in case of $f_{2,i}$. A related possibility is to "read" separately *blocks* of digits 0 and 1. Indeed, this has been done in [**2**] and [**4**] with a different aim: one is not interested in infinite sequences but rather in the set of words $w, f(w), f(f(w)), \dots$. Observe also that the way our sequences are formed is

somewhat similar to the growing of some of the sequences called *self-generating* in [12] (entries with numbers 21, 36, 70, 71, 89, 91, 201, 231, 254, 416, 425, 909, and 965.) The closest to our sequences is the *Kolakoski sequence* (nr. 70 in [12]). However, we take into account the whole previously generated word, not only a part of it (in fact, a symbol only in the Kalakoski case), when passing from a step to another one.

A natural generalization of the sequences in [11] is to consider an arbitrary base b and mappings [6]

$$f_{b,i}(x) = \beta_b(|x|_i)i, \qquad f'_{b,i}(x) = \beta_b(|x|_i),$$

for $i \in V_b = \{0,1,\dots,(b-1)\}$ ($\beta_b(n)$ is the representation of n in base b).

A further step is to count words, not necessarily digits: for $x, u \in V^+$, define [6]:

$$|x|_u = \mathrm{card}\{y \in V^* \mid x = yuy', y' \in V^*\},$$

and

$$f_{b,u}(x) = \beta_b(|x|_u)u, \qquad f'_{b,u}(x) = \beta_b(|x|_u).$$

Finally, we can "read" the length, following [7]:

$$f_{b,len}(x) = \beta_b(|x|).$$

Three problems were mainly investigated for these (parametrized) sequences $\alpha(x) \in V^\omega$: *ultimate periodicity* (is $\alpha(x)$ of the form $uvvv\dots$, for $u \in V^*$, $v \in V^+$?), *disjunctivity* (is every word $u \in V^+$ a subword of $\alpha(x)$?), and *ultimate identity* (given $\alpha(x)$, $\alpha(x')$, can we find $u_1, u_2 \in V^*$ and $\pi \in V^\omega$ such that $\alpha(x) = u_1\pi$, $\alpha(x') = u_2\pi$?).

The next table summarizes the results in [11], [10], [8], [9], [6], [7].

	ultimate periodicity	disjunctivity	ultimate identity
$z(f_{2,0}, x)$, $x \in V_2^+$	No	Yes	Yes
$z(f_{2,1}, x)$, $x \in V_2^+$	No	Yes	?
$z(f_{b,i}, x)$, $b \geq 3$, $i \in V_b - \{b-1\}$, $x \in V_b^+$	No	Yes	Yes
$z(f_{b,b-1}, x)$, $b \geq 3$, $x \in V_b^+$	No	Yes	?
$z(f'_{2,0}, x)$, $x \in V_2^+$	Yes	No	Yes
$z(f'_{2,1}, x)$, $x \in V_2^+$, $\lvert x\rvert_1 \geq 1$	No	Yes	No
$z(f'_{b,i}, x)$, $b \geq 3$, $i \in V_b$, $x \in V_b^+$	Yes	No	No
$z(f_{b,u}, x)$, $b \geq 2$, $u, x \in V_b^+$	No	Yes	?
$z(f'_{b,u}, x)$, $b \geq 2$, $u, x \in V_b^+$	Yes	No	No
$z(f_{b,len}, x)$, $b \geq 2$, $x \in V_b^+$	No	Yes	No

The basic tool for obtaining disjunctivity is the following

Lemma 1. *If $a_0 < a_1 < a_2 < \cdots$ are natural numbers such that $\lim_{n\to\infty}(a_{n+1}/a_n) = 1$, then for every $b \geq 2$ the sequence $\beta_b(a_0)\beta_b(a_1)\beta_b(a_2)\dots$ is disjunctive.*

For instance, in the case of $z(f_{b,i}, x) = x_0x_1x_2\dots$, with $x_0 = x$, and $x_{j+1} = f_{b,i}(x_0x_1\dots x_j) = \beta_b(|x_0\dots x_j|_i)i$, $j \geq 0$, we can take a_m to be the integer with the property $\beta_b(a_m) = x_m$, $m \geq 0$.

A similarly general lemma is used in [8], [6] for proving that a sequence is not ultimately periodic (Lemma 2 below); the other entries in the table are proved by ad-hoc combinatorial arguments.

For $b \geq 2$ and $x \in V_b^+$ define $\beta_b^{-1}(x) = n$ iff $x = 0^i\beta_b(n)$, $i \geq 0$.

Lemma 2. *If $b \geq 2$ and $f : V_b^+ \to B_b^+$ fulfills the following three conditions, then the sequence $z(f, w)$ is ultimately periodic for no $w \in V_b^+$:*

$$\beta_b^{-1}(f(xf(x))) > \beta_b^{-1}(f(x)), \text{ for all } x \in V_b^+, \tag{1}$$

$$\beta_b^{-1}(f(xy)) \leq \beta_b^{-1}(f(x)) + \beta_b^{-1}(f(y)), \text{ for all } x, y \in V_b^+, \tag{2}$$

$$\lim_{n \to \infty} \frac{\beta_b^{-1}(f(\beta_b(n)))}{n} = 0. \tag{3}$$

Note in the previous table the three *open problems* concerning ultimate identity.

A topic not yet investigated for self-reading sequences concerns *statistics of symbols* (and of *words*): Denote by $\alpha(n)$ the prefix of length n of a sequence $\alpha \in V^\omega$. What can one say about

$$\lim_{n \to \infty} \frac{|\alpha(n)|_i}{|\alpha(n)|_j},$$

for two symbols $i, j \in V$? The same question can be asked for i, j being words in V^+. In particular, are the self-reading sequences *Borel normal*? (Do any two words of equal length appear with the same probability in a given sequence?)

Of course, many other problems can be formulated (the reader might consult **[5]** and its references). For instance, have the self-reading sequences any *substitution property*, in the sense of **[1]**? (Which sequences are obtained by systematically replacing certain subwords by other words?)

REFERENCES

1. T. C. Brown, A characterization of the quadratic irrationals, *Canad. Math. Bull.*, 34 (1991) 36–41.
2. J. H. Conway, The weird and wonderful chemistry of audioactive decay, in *Open Problems in Communications and Computation* (T. M. Cover, B. Gopinath, eds.), Springer-Verlag, Berlin, 1987, 173–188.
3. J. Dassow, S. Marcus, Gh. Păun, Iterative reading of numbers and "black-holes", *Periodica Mathematica Hungarica*, 27, 2 (1993), 137–152.
4. J. Dassow, S. Marcus, Gh. Păun, Iterative reading of numbers: the ordered case, in *Developments in Language Theory* (G. Rozenberg, A. Salomaa, eds.), World Sci. Publ., Singapore, 1994, 157–168.
5. R. K. Guy, *Unsolved Problems in Number Theory*, Springer-Verlag, Berlin, 1981.
6. L. Ilie, On disjunctivity, ultimate periodicity and ultimate identity of Păun-Salomaa self-reading sequences, *Proc. of Developments in Language Theory* Conference, Magdeburg, July 1995.
7. L. Ilie, A new type of self-reading sequences, submitted, 1994.
8. G. Istrate, A note on self-reading sequences, *Discrete Applied Math.*, 50 (1994), 201–203.
9. G. Istrate, Gh. Păun, Some combinatorial properties of self-reading sequences, *Discrete Applied Math.*, 55 (1994), 83–86.
10. V. Mitrana, Some remarks on Păun-Salomaa sequences, *Bulletin of EATCS*, 49 (1993), 175–177.
11. Gh. Păun, A. Salomaa, Remarks concerning self-reading sequences, *Bulletin of EATCS*, 50 (1993) 229–233.
12. N. J. A. Sloane, *A Handbook of Integer Sequences*, Academic Press, New York & London, 1973. [The sequences referred to are listed in N. J. A. Sloane & S. Plouffe, *The Encyclopedia of Integer Sequences*, Academic Press, 1995, as M0068, 99, 190, 193, 250, 257, 557, 634, 689, 2633, 1112, 2303, 2438.]

Institute of Mathematics
Romanian Academy
P.O. Box 1-764
București 70700
Romania

Academy of Finland and
University of Turku
Department of Mathematics
20500 Turku
Finland

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS *address given on the inside front cover. Please include solutions and relevant references. Three copies of all items needed to evaluate the problem should be sent.*

Solutions of published problems should arrive at the MONTHLY PROBLEMS *address given on the inside front cover before July 31, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.*

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10501. *Proposed by Roger B. Eggleton, Illinois State University, Normal, IL.*

A *positive unitary fraction* is any rational number of the form $1/n$ where n is a positive integer. In how many ways can $1/n$ be expressed as

(**a**) the sum of two positive unitary fractions, or

(**b**) the difference of two positive unitary fractions?

10502. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.*

Let n and p be positive integers satisfying $1 \leq p \leq n$. Consider the permutation $\pi =$

$$\begin{pmatrix} 1 & 2 & \cdots & n-p & n-p+1 & n-p+2 & n-p+3 & \cdots & n \\ p+1 & p+2 & \cdots & n & p & 1 & 2 & \cdots & p-1 \end{pmatrix}.$$

Determine the cycle structure of π.

10503. *Proposed by Dragomir Ž. Đoković, University of Waterloo, Waterloo, Ontario, Canada.*

Show that

$$\left\{x^2 + (x+1)^2 + y^2 + z^2 : x, y, z \in \mathbb{Z}\right\}$$

is the set of positive integers not divisible by 4.

10504. *Proposed by Richard Hamming, Naval Postgraduate School, Monterey, CA, and Roger Pinkham, Stevens Institute of Technology, Hoboken, NJ.*

An urn contains a amber beads and b black beads with a and b both greater than zero. A bead is selected at random. If it is black, sampling stops; otherwise, it is replaced, an additional amber bead is added, and the process is repeated. Let N be the number of steps until the process stops.

(**a**) Show that $\mathbf{E}(N)$ is finite if $b > 1$ and find its value.

(**b**) Show that $\mathbf{E}(N)$ is infinite if $b = 1$.

(**c**) If n trials with $b = 1$ are performed, and $N_1, N_2, \ldots, N_n$ are the numbers of steps to completion in these trials, and $\overline{N}$ is their average, show that

$$Pr\left\{\left|\frac{\overline{N}}{\ln n} - 1\right| > \varepsilon\right\} \to 0$$

as $n \to \infty$.

10505. *Proposed by Fu-Chuen Chang, National Sun Yat-sen University, Kaohsiung, Taiwan.*

For $a, b \in \mathbb{R}$ with $a < b$ and $n \in \mathbb{Z}$ with $n \geq 2$ let

$$S_n(a,b) = \left\{(m_1, m_2) : m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2\right\},$$

with $\{x_1, \ldots, x_n\}$ ranging over all possible samples of n numbers in the interval $[a, b]$. Find the area $\Delta_n(a, b)$ of $S_n(a, b)$.

10506. *Proposed by David Callan, University of Wisconsin, Madison, WI.*

Let r be a positive integer and let Π_r denote the lattice of partitions of $\{1, 2, \ldots, r\}$ ordered by refinement, with typical element $\pi = \{\pi_1, \pi_2, \ldots, \pi_t\}$ where t is the number of blocks in π. Let μ denote the Möbius function for Π_r and let k and q be arbitrary. Show that

$$\sum_{\pi \in \Pi_r} \mu\left(\hat{0}, \pi\right)\left(|\pi_1|^k + |\pi_2|^k + \ldots + |\pi_t|^k\right)q^t = r!q\sum_{i=1}^{r}(-1)^{i+1}\binom{q}{r-i} i^{k-1}.$$

10507. *Proposed by Hauke Reddmann, University of Hamburg, Hamburg, Germany.*

Consider the equation

$$[0, a_1, a_2, \ldots, a_n] = 0.a_1a_2\cdots a_n \qquad (*)$$

where the expression on the left is a continued fraction and the expression on the right is a base b expansion. Assume that the a_i are all integers with $1 \leq a_i \leq b-1$.

(**a**) Find all solutions to $(*)$ when $n = 1$.

(**b**) Find all solutions to $(*)$ when $n = 2$.

(**c**)* Prove or disprove: there are no solutions when $n \geq 3$.

NOTES

(10506) Here, $\hat{0}$ denotes the minimal element of the lattice, which is the partition into singletons $\{\{1\}, \{2\}, \ldots, \{r\}\}$. Also, the Möbius function for this lattice is known: $\mu\left(\hat{0}, \pi\right) = (-1)^{r-t} \prod_i \left(|\pi_i| - 1\right)!$. Further details may be found in R. Stanley, *Enumerative Combinatorics*, Vol. 1. In particular, the case $k = 1$ is a consequence of Eq. (31) on p. 128 of that book. **(10507)** A finite continued fraction $[a_0, a_1, a_2, \ldots, a_n]$ represents a rational number. It is conventional to require that a_i be positive integers for $i > 0$, but one allows a_0 to be an arbitrary integer, assuring that all rational numbers can be represented. In particular, $a_0 = 0$ corresponds to the interval [0, 1]. The number represented is defined by induction on n: $[a_0]$ represents the integer a_0, and $[a_0, a_1, a_2, \ldots, a_n] = a_0 + 1/[a_1, a_2, \ldots, a_n]$. Note that $[a] = [a-1, 1]$, giving two representations of integers. Indeed, this construction extends to show that every rational number has two representations. No attempt will be made to choose between these representations to obtain a uniqueness result. However, restrictions have been put on the a_i that respect the usual restrictions on both representations of numbers.

SOLUTIONS

Asymptotics Via Bessel Functions

6675 [1991,965]. *Proposed by Alain Tissier, Montfermeil, France.*

Define

$$f(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^{1/n}}{n} \tag{0}$$

for $x > 0$. Determine the asymptotic behavior of $f(x)$ as $x \to 0$.

Solution by N. M. Temme, CWI, Amsterdam, The Netherlands. The solution relies on well-known properties of Bessel functions. These may be found in M. Abromowitz & I. A. Stegun, *Handbook of Mathematical Functions with Formulas Graphs and Mathematical Tables*, Nat. Bur. Standards Appl. Series 55, U. S. Government Printing Office, Washington, 1964, which will be denoted [**AS**] when cited below. Introduce α to stand for $\ln(1/x)$ to simplify the appearance of the various formulas for $f(x)$ in this solution. We first show that

$$f(x) = \int_0^{\infty} J_0\left(2\sqrt{\alpha t}\right) \frac{dt}{e^t + 1} \tag{1}$$

with $J_0(z)$ the Bessel function of order 0. Using the partial sums of the geometric series,

$$\frac{1}{e^t + 1} = e^{-t} \left(\sum_{n=0}^{N-1} (-1)^n e^{-nt} + (-1)^N \frac{e^{-Nt}}{1 + e^{-t}} \right)$$

for $N = 1, 2, \ldots$, and using a Laplace transform of the Bessel function [**AS**, p. 1026, formula 29.3.80], we write the right side of (1) in the form

$$\sum_{n=1}^{N-1} \frac{(-1)^{n-1} x^{1/n}}{n} + R_n(x)$$

where

$$R_N(x) = (-1)^N \int_0^\infty J_0\left(2\sqrt{\alpha t}\right) \frac{e^{-Nt}}{e^t+1}\, dt.$$

Since $|J_0(z)| \le 1$ for $z \ge 0$, we have $|R_N(x)| \le 1/(2N)$ for $N = 1, 2, \ldots$ and $0 < x \le 1$. This shows that the series in (0) converges to the function defined by (1).

Now, introduce the Hankel function $H_0^{(1)}(z)$ on the upper half plane, $0 < \arg(z) < \pi$, and analytically continue to the real axis. Formulas in chapter 9 of [**AS**] show that

$$H_0^{(1)}(-\bar{z}) = -\overline{H_0^{(1)}(z)},$$

allowing the integral in (1) to be rewritten to give

$$f(x) = \int_{-\infty}^{\infty} H_0^{(1)}\left(2u\sqrt{\alpha}\right) \frac{u\, du}{e^{u^2}+1}. \tag{2}$$

Next, we consider the asymptotic behavior of $f(x)$ as $x \to 0$. In fact, we derive a convergent expansion that has an asymptotic character. We use a standard technique from residue calculus and replace (2) by a series of residues due to the poles of $1 \big/ \left(\exp(u^2)+1\right)$ in the upper half plane. The poles are located at $\sqrt{(2k+1)\pi}\, e^{3\pi i/4}$ and $\sqrt{(2k+1)\pi}\, e^{\pi i/4}$ for $k = 0, 1, 2, \ldots$.

From section 9.9 of [**AS**] we find that

$$H_0^{(1)}\left(xe^{3i\pi/4}\right) = \frac{2}{\pi}\left(\operatorname{kei}(x) - i \operatorname{ker}(x)\right),$$

so that the residue series can be written as

$$f(x) = 4\sum_{k=0}^{\infty} \operatorname{ker}\left(2\sqrt{\alpha\pi(2k+1)}\right), \tag{3}$$

which converges if $0 < x \le 1$.

Using the asymptotic estimate

$$\operatorname{ker}(u) = \sqrt{\frac{\pi}{2u}} e^{-u/\sqrt{2}}\left(\cos\left(\frac{x}{\sqrt{2}} + \frac{\pi}{8}\right) + O\left(\frac{1}{x}\right)\right)$$

for $u \to \infty$ [**AS**, p. 381] we see that the sum in (3) over $k \ge 1$ is $O\left(\alpha^{-1/4} e^{-\sqrt{6\alpha\pi}}\right)$ while the term for $k = 0$ is $2(\pi/\alpha)^{1/4} e^{-\sqrt{2\alpha\pi}}\left(\cos(\sqrt{2\alpha\pi} + \pi/8) + O(\alpha^{-1/2})\right)$.

Recalling that $\alpha = -\ln x$, this gives

$$\begin{aligned} f(x) = {} & 2\left(\frac{-\pi}{\ln x}\right)^{1/4} e^{-\sqrt{-2\pi \ln x}} \cos\left(\sqrt{-2\pi \ln x} + \pi/8\right) \\ & + O\left((-\ln x)^{-3/4} e^{-\sqrt{-2\pi \ln x}}\right) \end{aligned}$$

as $x \to 0^+$. Alternatively, one could write $\alpha = t^2$ to obtain

$$f\left(e^{-t^2}\right) = 2\pi^{1/4} t^{-1/2} e^{-t\sqrt{2\pi}} \cos\left(t\sqrt{2\pi} + \pi/8\right) + O\left(t^{-3/2} e^{-t\sqrt{2\pi}}\right)$$

as $t \to \infty$.

Editorial comment. The contours used in the application of the residue theorem must avoid the singularities of the integrand. Albert Stadler provided more details on the construction of such contours and bounds on the integrals along them. Also note that the main term in this expression has infinitely many zeros as $x \to 0^+$. Near such values of x, the error term serves to isolate the zeros of the oscillatory function $f(x)$.

Solved also by D. Kucerovsky (student, U. K.), A. Stadler (Switzerland), and the proposer.

A q-trigonometric Identity

10226 [1992, 462]. *Proposed by Chu Wenchang, Academica Sinica, Beijing, China.*

Consider the functional equation

$$f(a-b)f(a-c)f(a-d)f(a-e) - f(b)f(c)f(d)f(e)$$
$$= q^b f(a)f(a-b-c)f(a-b-d)f(a-b-e)$$

with parameter q, where the variables a, b, c, d and e are related by

$$b+c+d+e = 2a.$$

(**a**) When $q = 1$, show that

$$f(\alpha) = \sin(k\alpha),$$

for any k, is a solution.

(**b**) When $0 < q < 1$, show that

$$f(\alpha) = (\Gamma_q(\alpha)\Gamma_q(1-\alpha))^{-1}$$

is a solution.

Solution of (a) by Shalosh B. Ekhad, Temple University, Philadelphia, PA. Computer algebra systems have made this routine. For example, the following procedure in *Maple* verifies the identity by returning a value of zero:

```
q:=1;
f:=proc(al):
(z^al-z^(-al))/2/I:
end:
b:=2*a-c-d-e:
F:=f(a-b)*f(a-c)*f(a-d)*f(a-e)-f(b)*f(c)*f(d)*f(e)-
q^b*f(a)*f(a-b-c)*f(a-b-d)*f(a-b-e);
simplify(F);
```

Here z is used for e^{ik}.

Solution of (b) by Robin J. Chapman, University of Exeter, Exeter, U. K.. Consider the theta functions

$$\vartheta_1(z) = 2q^{1/8}\sum_{n=0}^{\infty}(-1)^n q^{n(n+1)/2}\sin(2n+1)z,$$

$$\vartheta_2(z) = 2q^{1/8}\sum_{n=0}^{\infty} q^{n(n+1)/2}\cos(2n+1)z,$$

$$\vartheta_3(z) = 1+2\sum_{n=1}^{\infty} q^{n^2/2}\cos 2nz$$

and

$$\vartheta_4(z) = 1+2\sum_{n=1}^{\infty}(-1)^n q^{n^2/2}\cos 2nz.$$

(The dependence on the parameter q is slightly different in [2].) These are entire functions of z, with ϑ_1 being odd, and the rest even.

A classical identity of Jacobi [2, §21.22] states that

$$\begin{aligned}
&2\vartheta_1\left(\frac{-b+c+d+e}{2}\right)\vartheta_1\left(\frac{b-c+d+e}{2}\right)\\
&\quad\times\vartheta_1\left(\frac{b+c-d+e}{2}\right)\vartheta_1\left(\frac{b+c+d-e}{2}\right)\\
&=\vartheta_1(b)\vartheta_1(c)\vartheta_1(d)\vartheta_1(e)+\vartheta_2(b)\vartheta_2(c)\vartheta_2(d)\vartheta_2(e)\\
&\quad-\vartheta_3(b)\vartheta_3(c)\vartheta_3(d)\vartheta_3(e)+\vartheta_4(b)\vartheta_4(c)\vartheta_4(d)\vartheta_4(e).
\end{aligned}$$

Replacing b by $-b$ and using the parities of the theta functions gives

$$\begin{aligned}
&2\vartheta_1\left(\frac{b+c+d+e}{2}\right)\vartheta_1\left(\frac{-b-c+d+e}{2}\right)\\
&\quad\times\vartheta_1\left(\frac{-b+c-d+e}{2}\right)\vartheta_1\left(\frac{-b+c+d-e}{2}\right)\\
&=-\vartheta_1(b)\vartheta_1(c)\vartheta_1(d)\vartheta_1(e)+\vartheta_2(b)\vartheta_2(c)\vartheta_2(d)\vartheta_2(e)\\
&\quad-\vartheta_3(b)\vartheta_3(c)\vartheta_3(d)\vartheta_3(e)+\vartheta_4(b)\vartheta_4(c)\vartheta_4(d)\vartheta_4(e).
\end{aligned}$$

Comparing these two identities we see that

$$\begin{aligned}
&\vartheta_1(a-b)\vartheta_1(a-c)\vartheta_1(a-d)\vartheta_1(a-e)-\vartheta_1(b)\vartheta_1(c)\vartheta_1(d)\vartheta_1(e)\\
&\quad=\vartheta_1(a)\vartheta_1(a-b-c)\vartheta_1(a-b-d)\vartheta_1(a-b-e).
\end{aligned}$$

Another classical result [2, §21.3] gives the theta functions as infinite products. In particular

$$\vartheta_1(z)=2A_1(q)\sin z\prod_{n=1}^{\infty}(1-q^n\exp 2iz)(1-q^n\exp -2iz)$$

where $A_1(q)$ is a function of q. Hence, recalling the definition of the q-gamma function,

$$\Gamma_q(\alpha)=\frac{(q;q)_\infty}{(q^\alpha;q)_\infty}(1-q)^{1-\alpha}$$

where

$$(x;q)_\infty=\prod_{n=0}^{\infty}(1-xq^n),$$

we have

$$\begin{aligned}
f(\alpha)&=A_2(q)\prod_{n=0}^{\infty}(1-q^nq^\alpha)(1-q^{n+1}q^{-\alpha})\\
&=A_3(q)q^{\alpha/2}(q^{\alpha/2}-q^{-\alpha/2})\prod_{n=1}^{\infty}(1-q^nq^\alpha)(1-q^nq^{-\alpha})\\
&=A_4(q)q^{\alpha/2}\vartheta_1(i(\alpha/2)\log q)
\end{aligned}$$

where $A_2(q)$, $A_3(q)$, and $A_4(q)$ are functions of q. Using this expression for $f(\alpha)$ in our theta function identity gives

$$\begin{aligned}
&q^{(b+c+d+e-4a)/2}f(a-b)f(a-c)f(a-d)f(a-e)\\
&\quad-q^{-(b+c+d+e)/2}f(b)f(c)f(d)f(e)\\
&=q^{(3b+c+d+e-4a)/2}f(a)f(a-b-c)f(a-b-d)f(a-b-e)
\end{aligned}$$

or

$$\begin{aligned}
&f(a-b)f(a-c)f(a-d)f(a-e)-f(b)f(c)f(d)f(e)\\
&\quad=q^bf(a)f(a-b-c)f(a-b-d)f(a-b-e)
\end{aligned}$$

as required. Note that the particular form of $A_4(q)$ was not used.

Editorial comment. The proposer's proof of part **(b)** was based on writing $p = q$ in [**1**, eq. 3.6.14, p. 72]. This, in turn, is based on using formulas for differences of expressions involving $(z, q)_n = \prod_{k=0}^{n-1}(1 - zq^k)$ to telescope series involving these quantities. Problem 2.16 of [**1**] is also equivalent to **(a)** under a change of variables. This theory also includes a derivation of the expressions giving the theta functions as infinite products. Incorporation of these formulas in symbolic packages may yet make results like part **(b)** routine.

The q-gamma function is defined by $\Gamma_q(\alpha) = \dfrac{(q;q)_\infty}{(q^\alpha;q)_\infty}(1-q)^{1-\alpha}$ where $(x;q)_\infty = \prod_{n=0}^{\infty}(1 - xq^n)$. At any rate, once one has **(b)** one can take limits as $q \to 1^-$. The function $\Gamma_q(z)$ approaches the classical Gamma function [**1**, section 1.10]. Since $\Gamma(\alpha)\Gamma(1-\alpha) = \pi/\sin\pi\alpha$ [**2**, Section 12.14, p. 239], a multiple of $\sin\pi\alpha$ satisfies the equation with $q = 1$. As has been noted, homogeneity then shows that $f(\alpha) = C\sin(k\alpha)$ satisfies the equation for all C and k.

Robin Chapman gave the full details of the translation of the identity of part **(a)** into complex exponentials. When a is eliminated, each term in the formula is symmetric in b, c, d and e. The resulting expressions were easy to work with although their display took more space than other solutions.

Miklós Mócsy solved part **(a)** by fixing a, b and c and considering the difference of the two sides of the desired identity as $G(d)$. It is easy to verify that $G(0) = 0$. In terms of the quantity s defined by $2s = 2a - b - c$, the addition formulas of trigonometry allow a factor of $\cos(s) - \cos(s - d)$ to be identified in $G(d) - G(0)$. The complementary factor is then easily simplified by the addition formulas to zero.

REFERENCES

1. G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge, 1990.
2. E. T. Whittaker & G. N. Watson, *A Course of Modern Analysis* Fourth Edition, Cambridge, 1927.

Part **(a)** solved also by M. Mócsy (Hungary). The proposer also solved part **(b)**, and noted that part **(a)** follows.

Polynomials With No Zeros in a Disk

10255 [1992, 782]. *Proposed by Zalman Rubinstein, University of Haifa, Haifa, Israel.*

Let $P_n(z)$ be a polynomial of degree n having no roots in the open unit disk

$$\mathcal{D} = \{z : |z|| < 1\}.$$

(a) For all real η, show that the polynomial

$$P_n(z) - \left(1 - e^{i\eta}\right)\frac{zP_n'(z)}{n}$$

also has no roots in $\mathcal{D}$.

(b) For $0 < p < 1$, show that

$$\int_0^{2\pi}\left|P'\left(e^{it}\right)\right|^p dt \le C_p n^p \int_0^{2\pi}\left|P\left(e^{it}\right)\right|^p dt,$$

with

$$C_p = \frac{2\pi}{\int_0^{2\pi}\left|1 + e^{it}\right|^p dt} = \frac{2^{-p}\pi^{\frac{1}{2}}\Gamma\left(\frac{1}{2}p + 1\right)}{\Gamma\left(\frac{1}{2}p + \frac{1}{2}\right)}.$$

(c) Determine all polynomials for which the inequality in **(b)** becomes an equality.

Solution by Antonios D. Melas, Athens, Greece. The three parts will be proved in order.

(**a**) We may assume $e^{i\eta} \neq 1$. We can write $P_n(z) = A(z-a_1)\dots(z-a_n)$ where $|a_j| \geq 1$ for $1 \leq j \leq n$. Let $b_j = 1/a_j$ (so $|b_j| \leq 1$). Then for $|z| < 1$ we have

$$\begin{aligned}
\frac{zP_n'(z)}{P_n(z)} - \frac{n}{1-e^{i\eta}} &= \sum_{j=1}^{n}\left(\frac{z}{z-a_j}\right) - \frac{n}{1-e^{i\eta}} \\
&= \sum_{j=1}^{n}\left(\frac{b_j z}{b_j z - 1}\right) - \frac{n}{1-e^{i\eta}} \\
&= \frac{1}{2}\sum_{j=1}^{n}\left(1 - \frac{1+b_j z}{1-b_j z}\right) - \frac{n}{1-e^{i\eta}} \\
&= -\frac{1}{2}\left(\sum_{j=1}^{n}\left(\frac{1+b_j z}{1-b_j z}\right) - n + \frac{2n}{1-e^{i\eta}}\right) \\
&= -\frac{1}{2}\left(\sum_{j=1}^{n}\left(\frac{1+b_j z}{1-b_j z}\right) + n\frac{1+e^{i\eta}}{1-e^{i\eta}}\right).
\end{aligned}$$

Since $|b_j z| < 1$ we have $\Re\frac{1+b_j z}{1-b_j z} > 0$ for $1 \leq j \leq n$. Also $\Re\frac{1+e^{i\eta}}{1-e^{i\eta}} = 0$. Hence $\Re\left(\frac{zP_n'(z)}{P_n(z)} - \frac{n}{1-e^{i\eta}}\right) < 0$, and $P_n(z) - \left(1-e^{i\eta}\right)\frac{zP_n'(z)}{n}$ has no roots in $\mathfrak{D}$.

(**b**) It is helpful to formulate three lemmas before beginning the main part of the proof.

Lemma 1. *If $P(z)$ is a polynomial of degree n with no roots in the unit disc $\mathfrak{D}$ and if η is a real number then:*

$$\int_0^{2\pi} \log\left|P\left(e^{it}\right) - \left(1-e^{i\eta}\right)\frac{e^{it}P'\left(e^{it}\right)}{n}\right| dt = \int_0^{2\pi} \log\left|P\left(e^{it}\right)\right| dt. \tag{1}$$

Proof. Let $Q(z) = P(z) - \left(1-e^{i\eta}\right)\frac{zP'(z)}{n}$. Then by the assumption and by (**a**) both $P(z)$ and $Q(z)$ have no roots in $\mathfrak{D}$. Hence by Jensen's formula we have:

$$\int_0^{2\pi} \log\left|P\left(e^{it}\right)\right| dt = \log|P(0)|$$

$$\int_0^{2\pi} \log\left|Q\left(e^{it}\right)\right| dt = \log|Q(0)|.$$

Since $P(0) = Q(0)$, equation (1) follows.

Lemma 2. *If $P(z)$ is any polynomial of degree n and if η is a real number then:*

$$\int_0^{2\pi} \log\left|P\left(e^{it}\right) - \left(1-e^{i\eta}\right)\frac{e^{it}P'\left(e^{it}\right)}{n}\right| dt \leq \int_0^{2\pi} \log\left|P\left(e^{it}\right)\right| dt. \tag{2}$$

Proof. Let $a_1, \dots, a_n$ be the roots of $P(z)$ with $|a_j| < 1$ for $1 \leq j \leq k$ and $|a_j| \geq 1$ for $k < j \leq n$. Consider the function

$$F(\zeta_1, \dots, \zeta_k, t) = \log\left|1 - \frac{1-e^{i\eta}}{n}\left(\sum_{j=1}^{k}\frac{e^{it}}{e^{it}-\zeta_j} + \sum_{j=k+1}^{n}\frac{e^{it}}{e^{it}-a_j}\right)\right|.$$

For every t with e^{it} not equal to any of $a_{k+1}, \dots, a_n$, F is the logarithm of the absolute value of an analytic function in $\zeta_1, \dots, \zeta_k$; and hence, a subharmonic function of $\zeta_1, \dots, \zeta_k$

when $|\zeta_j| < 1$ for $1 \le j \le k$. Using the characterization of subharmonic functions with submeanvalue inequalities and Fubini's theorem we conclude that the function:

$$F(\zeta_1, \dots, \zeta_k) = \int_0^{2\pi} F(\zeta_1, \dots, \zeta_k, t)\, dt$$

is subharmonic for $|\zeta_j| < 1$ $(1 \le j \le k)$. (Fubini's theorem requires

$$\int_0^{2\pi} |F(\zeta_1, \dots, \zeta_k, t)|\, dt < \infty,$$

but that is easily verified.) By the maximum principle for subharmonic functions there exist $c_1, \dots, c_k$ with $|c_j| = 1$ $(1 \le j \le k)$ such that

$$F(a_1, \dots, a_k) \le F(c_1, \dots, c_k). \tag{3}$$

Introduce the polynomial $\tilde{P}(z) = P(z) \prod_{j=1}^{k} \frac{z - c_j}{z - a_j}$, that now has no roots in $\mathcal{D}$. Then, by Lemma 1, we have:

$$\begin{aligned} F(c_1, \dots, c_k) &= \int_0^{2\pi} \log \left| \frac{\tilde{P}(e^{it}) - (1 - e^{i\eta}) \frac{e^{it}\tilde{P}'(e^{it})}{n}}{\tilde{P}(e^{it})} \right| dt \\ &= 0. \end{aligned}$$

Hence (3) implies (2).

Lemma 3. *For the polynomial $P_n(z)$ and for any real number η (and using $\log^+ x$ for $\max(0, \log x)$) we have:*

$$\int_0^{2\pi} \log^+ \left| P_n\left(e^{it}\right) - \left(1 - e^{i\eta}\right) \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right| dt \le \int_0^{2\pi} \log^+ \left| P_n\left(e^{it}\right) \right| dt. \tag{4}$$

Proof. Take $P(z) = P_n(z) + e^{i\theta}$ in Lemma 2 and then integrate for $0 \le \theta \le 2\pi$. Using the consequence of Jensen's formula:

$$\int_0^{2\pi} \log \left| z + e^{i\theta} \right| d\theta = 2\pi \log^+ |z|$$

for any $z \in \mathbb{C}$, we obtain (4).

This completes the lemmas.

Now using the equation:

$$x^p = p^2 \int_0^{\infty} \log^+ \left(\frac{x}{s} \right) s^{p-1} ds$$

for $x > 0$ and $p > 0$, and introducing $Q_n(z) = P_n(z) - \left(1 - e^{i\eta}\right) \frac{z P_n'(z)}{n}$, we have

$$\begin{aligned} \int_0^{2\pi} \left| Q_n\left(e^{it}\right) \right|^p dt &= p^2 \int_0^{\infty} \int_0^{2\pi} \log^+ \left| \frac{Q_n\left(e^{it}\right)}{s} \right| dt\, s^{p-1} ds \\ &\le p^2 \int_0^{\infty} \int_0^{2\pi} \log^+ \left| \frac{P_n\left(e^{it}\right)}{s} \right| dt\, s^{p-1} ds \\ &= \int_0^{2\pi} \left| P_n\left(e^{it}\right) \right|^p dt. \end{aligned}$$

Hence:

$$\int_0^{2\pi} \left| P_n\left(e^{it}\right) - \left(1 - e^{i\eta}\right) \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right|^p dt \le \int_0^{2\pi} \left| P_n\left(e^{it}\right) \right|^p dt \tag{5}$$

for $p > 0$.

For $p > 0$ and $a, b \in \mathbb{C}$ we have:

$$\int_0^{2\pi} \left| a + b e^{i\eta} \right|^p d\eta \ge \min\left(|a|^p, |b|^p \right) \int_0^{2\pi} \left| 1 + e^{i\eta} \right|^p d\eta. \tag{6}$$

It suffices to prove this for $b = 1, a > 1$ since the integral is invariant under the substitutions $\eta \mapsto -\eta$ and $\eta \mapsto \eta + \theta$. In this case, (6) follows from $\left|a + e^{i\eta}\right| \ge \left|1 + e^{i\eta}\right|$.

Integrating (5) over $0 \le \eta \le 2\pi$ and using (6) we get

$$\int_0^{2\pi} \min\left(\left| P_n\left(e^{it}\right) - \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right|^p, \left| \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right|^p \right) dt \le C_p \int_0^{2\pi} \left| P_n\left(e^{it}\right) \right|^p dt \tag{7}$$

for $p > 0$. We claim that

$$\left| \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right| \le \left| P_n\left(e^{it}\right) - \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right| \tag{8}$$

for all t real. The result is clear when e^{it} is a root of $P_n(z)$. If e^{it} is not a root of $P_n(z)$, divide (8) by $P_n\left(e^{it}\right)$ to get the equivalent condition

$$\left| \frac{e^{it} P_n'\left(e^{it}\right)}{n P_n\left(e^{it}\right)} \right| \le \left| 1 - \frac{e^{it} P_n'\left(e^{it}\right)}{n P_n\left(e^{it}\right)} \right|$$

This means that the distance from $\dfrac{e^{it} P_n'\left(e^{it}\right)}{n P_n\left(e^{it}\right)}$ to 1 is bigger than or equal to the distance from it to 0 hence is equivalent to

$$\Re \frac{e^{it} P_n'\left(e^{it}\right)}{n P_n\left(e^{it}\right)} \le \frac{1}{2}.$$

Writing $P_n(z) = A(z - a_1) \cdots (z - a_n)$ where $\left|a_j\right| \ge 1$ this is equivalent to:

$$\Re \sum_{j=1}^{n} \frac{e^{it}}{e^{it} - a_j} \le \frac{n}{2}$$

hence to

$$\Re \sum_{j=1}^{n} \frac{e^{it} + a_j}{e^{it} - a_j} \le 0.$$

This is true since $\left|a_j\right| \ge 1$ implies $\Re \dfrac{e^{it} + a_j}{e^{it} - a_j} \le 0$ if $e^{it} \ne a_j$. Hence (8) holds.

Combining (7) and (8) we get

$$\int_0^{2\pi} \left| \frac{e^{it} P_n'\left(e^{it}\right)}{n} \right|^p dt \le C_p \int_0^{2\pi} \left| P_n\left(e^{it}\right) \right|^p dt$$

hence

$$\int_0^{2\pi} \left| P_n'\left(e^{it}\right) \right|^p dt \le C_p n^p \int_0^{2\pi} \left| P_n\left(e^{it}\right) \right|^p dt \tag{9}$$

for all $p > 0$.

(c) If equality holds in (9) for $P_n(z)$ then equality must hold in (8) for every t and also equality must hold in (5) for any real η and therefore equality must hold in (4) with $P_n(z)$ replaced by $\frac{1}{s}P_n(z)$ and any $s > 0$.

From the proof of (8), equality in (8) requires that $\Re\dfrac{e^{it}+a_j}{e^{it}-a_j} = 0$ for every t with $P_n(e^{it}) \neq 0$ and every root a_j of $P_n(z)$, $1 \leq j \leq n$. Hence all roots of $P_n(z)$ are on the unit circle. Since equality holds in (4) for $\frac{1}{s}P_n(z)$ using $x^2 = 4\int_0^\infty \log^+(x/s)s\,ds$ we conclude that

$$\int_0^{2\pi} \left| P_n\left(e^{it}\right) - \left(1 - e^{i\eta}\right)\frac{e^{it}P_n'\left(e^{it}\right)}{n} \right|^2 dt = \int_0^{2\pi} \left|P_n\left(e^{it}\right)\right|^2 dt. \tag{10}$$

Writing $P_n(z) = a_n z^n + \cdots + a_j z + a_0$ and using Parseval's identity and (10) we get:

$$\sum_{j=0}^{n} \left| a_j \left[1 - \frac{j}{n}\left(1 - e^{i\eta}\right)\right]\right|^2 = \sum_{j=0}^{n} |a_j|^2$$

hence:

$$\sum_{j=0}^{n} 4\left(\sin^2\frac{\eta}{2}\right)\frac{j}{n}\left(1 - \frac{j}{n}\right)|a_j|^2 = 0$$

for every η, hence:

$$\sum_{j=0}^{n} j(n-j)\,|a_j|^2 = 0$$

and since $j(n-j) > 0$ for $1 \leq j \leq n-1$ we must have: $a_1 = \cdots = a_{n-1} = 0$ hence $P_n(z) = az^n + b$, $a, b \in \mathbb{C}$. Since also all the roots of $P_n(z)$ lie on the unit circle $\{z : |z| = 1\}$ we must have $|a| = |b| \neq 0$. Conversely if $P_n(z) = az^n + b$ where $|a| = |b| \neq 0$ it is easy to see that equality holds in (9). Hence equality holds in (9) if and only if $P_n(z) = az^n + b$ with $|a| = |b| \neq 0$.

Solved also by I. Kastanas and the proposer.

A Cubic Relative of the AGM

10281 [1993, 76]. *Proposed by Jonathan M. Borwein, University of Waterloo, Waterloo, Ontario, Canada.*

For $a > 0$ and $b > 0$, let

$$I(a,b) = \int_0^\infty \frac{t\,dt}{\sqrt[3]{(a^3+t^3)(b^3+t^3)^2}}.$$

(a) Show that

$$I(a,b) = I\left(\frac{a+2b}{3}, \sqrt[3]{b\frac{a^2+ab+b^2}{3}}\right).$$

(b) Show that the iteration which has $a_0 = a$ and $b_0 = b$ and

$$a_{n+1} = \frac{a_n + 2b_n}{3}$$
$$b_{n+1} = \sqrt[3]{b_n\frac{b_n^2 + a_n b_n + a_n^2}{3}}$$

converges to $I(1,1)/I(a,b)$.

Solution of **(a)** *by Donald A. Darling, Newport Beach, CA.* Let $f(x^3) = I(x, 1)$ so that a change of variables $t \mapsto bt$ yields $I(a, b) = (1/b) \cdot f(a^3/b^3)$. Let $x = a/b$, and write

$$a' = \frac{a+2b}{3} = b\frac{x+2}{3}$$
$$b' = \left(b\,\frac{b^2+ab+a^2}{3}\right)^{1/3} = b\left(\frac{1+x+x^2}{3}\right)^{1/3}.$$

In this notation, the proposed identity, $I(a, b) = I(a', b')$, becomes

$$f\left(x^3\right) = \left(\frac{3}{1+x+x^2}\right)^{1/3} f\left(1 - \left(\frac{(1-x)^3}{9(1+x+x^2)}\right)\right).$$

We use some transformations to recognize the function $f(x)$ and to simplify this expression. This follows chapter XIV of E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* Fourth Edition, Cambridge, 1927, especially section 14.53 relating hypergeometric functions of z and $1 - z$ and section 14.6 expressing integrals of the given type as hypergeometric functions. The change of variables $t = \sqrt[3]{\frac{1}{u} - 1}$ identifies $f(x)$ as $C \cdot {}_2F_1\left(1/3, 1/3; 1; (1-x)\right)$ with $C = \Gamma(1/3)\Gamma(2/3)/3$. If $g(x) = {}_2F_1(2/3, 1/3; 1; x)$, then known transformations give $f(x) = Cx^{-1/3}g\left(\frac{x-1}{x}\right)$. The function $g(x)$ satisfies the differential equation

$$x(1-x)g''(x) + (1-2x)g'(x) - {}^2\!/\!_9\,g(x) = 0$$

and has a power series expansion

$$g(x) = \sum_{k=0}^{\infty} \frac{(3k)!}{3^{3k}(k!)^3} x^k$$

convergent for $|x| < 1$. In terms of the function g, the proposed relation becomes

$$(2x+1)g\left(x^3\right) = g\left(1 - \left(\frac{1-x}{2x+1}\right)^3\right).$$

It can be verified, using the differential equation satisfied by g, that each side of this expression satisfies the differential equation

$$x(1+2x)^2(1-x^3)w''(x) + (1+2x)(1+2x^2)(1-2x-2x^2)w'(x) - 2(1-x^2)w(x) = 0.$$

The power series for g also shows that the power series for both sides begin with the terms

$$1 + 2x + \frac{2}{9}x^3 + \frac{10}{81}x^6 + \cdots.$$

Thus the equality is established.

Solution of **(b)** *by D. B. Tyler, Hughes Aircraft, El Segundo, CA.* First note that

$$a_{n+1}^3 - b_{n+1}^3 = \left(\frac{a_n - b_n}{3}\right)^3.$$

Thus, if $a \geq b > 0$, then $a_n \geq b_n > 0$ for all n. Similarly, if $0 < a \leq b$, then $0 < a_n \leq b_n$ for all n. Since a_{n+1}^3 and b_{n+1}^3 are weighted averages of a_n^3, $a_n^2 b_n$, $a_n b_n^2$ and b_n^3, it follows that a_{n+1} and b_{n+1} lie between a_n and b_n. Thus the sequences $\langle a_n \rangle$ and $\langle b_n \rangle$ are convergent, since they are monotone and bounded. The definition of the iteration then shows that they have a common limit, say L.

Since $\left(x^3+t^3\right)\left(y^3+t^3\right)^2$ lies between $\left(x^3+t^3\right)^3$ and $\left(y^3+t^3\right)^3$, it follows that $I(a,b)=I(a_n,b_n)$ must lie between $I(a_n,a_n)=I(1,1)/a_n$ and $I(b_n,b_n)=I(1,1)/b_n$ (where the scaling used in the solution of **(a)** has been used to show that $xI(x,x)=I(1,1)$). Letting $n\to\infty$ gives $I(a,b)=I(1,1)/L$, as desired.

Editorial comment. Yves Dumont noted that the classical AGM and $I(a,b)$ are the cases $p=2$ and $p=3$ of

$$I_p(a,b)=\int_0^\infty \frac{x^{p-2}\,dx}{(x^p+a^p)^{1/p}(x^p+b^p)^{(p-1)/p}}$$

that are related to solutions of the differential equation

$$x\left(1-x^p\right)Y''+\left(1-(p+1)x^p\right)Y'-(p-1)x^{p-1}Y=0. \qquad (E_p)$$

When $p=2$ or $p=3$, there is a modular transformation for the solutions of (E_p) that are bounded around zero. If $J_p(x)$ is one of these solutions, the transformation takes the form $J_p(\lambda)=\mu J_p(x)$ where $\lambda=\left(1-u\right)\,/\,\left(1+(p-1)u\right)$ and $\mu=\left(1+(p-1)u\right)\,/\,p$ with u given by $x^p+u^p=1$. Modular transformations are also known when $p=4$ and $p=6$, but an investigation of the power series shows that these do not give identities for the J_p.

The solution of part **(a)** by the proposer was similar to the selected solution. Reference was given to J. M. Borwein and P. B. Borwein, "A remarkable cubic iteration", *Computational Methods and Function Theory*, Springer Lecture Notes in Mathematics #1435, 1990, pp. 27–31 and J. M. Borwein and P. B. Borwein, "A cubic counterpart of Jacobi's identity and the AGM", *Trans. Amer. Math. Soc.* 323 (1991), 691–701. There is still no self-contained proof that avoids exploiting the identification with a hypergeometric function.

Solved also by Y. Dumont (France), A. D. Melas (Greece), and the proposer.

Special Inner Automorphisms

10301 [1993, 401]. *Proposed by William P. Wardlaw, United States Naval Academy, Annapolis MD.*

Let R be a commutative ring with identity. For which matrices A in $\mathbf{GL}_n(R)$ is the mapping

$$\alpha_A\colon \mathbf{SL}_n(R)\to\mathbf{SL}_n(R) \text{ defined by } X\mapsto AXA^{-1}$$

an *inner* automorphism of $\mathbf{SL}_n(R)$.

Solution by Richard Holzsager, The American University, Washington, DC. The condition holds if and only if $\det A$ is an nth power. If A is a matrix for which the mapping is inner, then there is a matrix B in $\mathbf{SL}_n(R)$ such that $AXA^{-1}=BXB^{-1}$ for all X is $\mathbf{SL}_n(R)$. In other words, $B^{-1}A$ commutes with every such X, including $X=I+E_{ij}$, where I is the identity matrix and E_{ij} is the matrix with 1 in position i,j and 0 elsewhere. Therefore, $B^{-1}A$ commutes with E_{ij}. This implies that the ith column and the jth row of $B^{-1}A$ are zero except for their diagonal elements, which agree. Varying i and j, we find that $B^{-1}A$ is a scalar multiple of I; equivalently, $A=rB$, which implies that $\det A=r^n$. This condition is also sufficient, as we can see by taking $B=r^{-1}A$, where we know that r is invertible because A is.

Solved also by S. F. Barger, M. Barile (student, Germany), V. Božin (student, Yugoslavia), D. Callan, R. J. Chapman (U. K.), G. Ehrlich, H. von Eitzen (Germany), F. J. Flanigan, S. M. Gagola Jr., I. Kastanas, D. W. Koster, J. J. Kuzmanovich, D. C. Lantz, F. C. Leary, J. H. Lindsey II, O. P. Lossers (The Netherlands), F. Schmidt, R. W. Sheets, A. N. 't Woord (The Netherlands), GCHQ Problem Solving Group (U. K.) (two solutions), and the proposer. Three incorrect solutions were received.

Mathematics has beauties of its own—a symmetry and proportion in its results, a lack of superfluity, an exact adaptation of means to ends, which is exceedingly remarkable and to be found only in the works of the greatest beauty... When this subject is properly... presented, the mental emotion should be that of enjoyment of beauty, not that of repulsion from the ugly and the unpleasant.

—*J. W. A. Young*

Mathematical Circles Squared, Howard W. Eves, Boston: Prindle, Weber and Schmidt.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

Introduction to Hyperbolic Geometry. By Arlan Ramsay and Robert D. Richtmyer. Springer-Verlag, New York, Inc., 1995, vi + 287, $39.00.

Reviewed by **John G. Ratcliffe**

The modern form of Euclid's parallel postulate says that through a point outside of a straight line there is a unique straight line passing through the point that is parallel to the given line. For over two thousand years geometers struggled with understanding the role of the parallel postulate in the logical structure of Euclidean geometry. Many believed that the parallel postulate could be proved from the other axioms of plane geometry and sought vainly for its proof. Not until the second decade of the nineteenth century did Gauss in Germany, Lobachevsky in Russia, and Bolyai in Hungary discover, independently, that the denial of the parallel postulate leads to a new strange geometry that Gauss called non-Euclidean geometry and which today is called hyperbolic geometry.

This new geometry was so revolutionary in its day that Gauss felt it was prudent not to publish anything about it. The burden of presenting hyperbolic geometry as a possible alternative to the monumental edifice of Euclidean geometry fell to Lobachevsky, who was severely criticized for his efforts. Gauss learned Russian so he could read the criticism. Bolyai's efforts did not bear much fruit either. When Bolyai learned that Gauss claimed prior knowledge of hyperbolic geometry, he felt cheated out of his discoveries and published nothing else on the subject.

Gauss, Lobachevsky, and Bolyai developed hyperbolic geometry from first principles in a tour de force of traditional synthetic geometry. They discovered many theorems; however, they did not give a convincing proof that hyperbolic geometry is consistent or at least as consistent as Euclidean geometry. This honor fell upon the Italian geometer Beltrami in 1868 when he presented a Euclidean model of hyperbolic geometry thereby proving that hyperbolic geometry is as consistent as Euclidean geometry. Beltrami used differential geometry to show that his model satisfies the axioms of hyperbolic geometry.

In the book under review, the authors lead you along essentially the same path as that followed by Gauss, Lobachevsky, Bolyai, and Beltrami. The first four chapters present an axiomatic development of hyperbolic geometry. In Chapters 5, 6, and 7, the differential geometry of surfaces is developed and used to prove the consistency of the set of axioms of hyperbolic geometry. Chapters 8 and 9 are concerned with properties of models of hyperbolic geometry in two and three dimensions. The relationship between hyperbolic geometry and the theory of special relativity is discussed in Chapter 10. The last chapter is about classical construction problems in the hyperbolic plane.

There are three main themes of this book. The first is the development of geometric intuition through synthetic geometry in Chapters 1–4; the second is the

appreciation of the logical structure of the subject in Chapters 7 and 11; and the third is the introduction of differential geometry in Chapters 5, 6, and 9. Each theme is skillfully developed and intertwined with the others into a coherent whole.

Some of the highlights of the book are the following: (1) The axioms of hyperbolic geometry are stated in a modern form. They include the axioms of a metric space. The axioms of hyperbolic geometry are not only shown to be consistent but also categorical, that is, they are shown to determine hyperbolic geometry completely and uniquely. (2) The trigonometric formulas of hyperbolic geometry are derived directly from the axioms, without use of a model. The locally Euclidean nature of the hyperbolic plane is established by showing that certain Euclidean laws are approximately satisfied by small figures, with a relative error that tends to zero as the size of the figure tends to zero. This analysis makes it possible to set up differential equations to derive the formulas of hyperbolic geometry. (3) The introduction to the differential geometry of surfaces in Chapter 5 is excellent. A nice feature for physics students is that the terminology of differential geometry is compared with the terminology used by physicists.

This is an excellent book. It is more than an introduction to the subject. It is an adventure in the realm of mathematics. The strange new world of hyperbolic geometry is introduced and explored. Geometry and analysis are synthesized to form differential geometry and used to tame this new world. Discussions of how hyperbolic geometry fits into the scheme of mathematics and physics adds excitement to the adventure.

Although the first four chapters together with selections from the rest of the book can serve as a textbook for a standard upper division course on hyperbolic geometry, the real strength of this book is as a textbook for a year long introduction to differential geometry for advanced undergraduates. I highly recommend this book to the readers of the Monthly who are interested in higher geometry.

Department of Mathematics
Vanderbilt University
1326 Stevenson Ctr.
Nashville, TN 37240-0001

Elements of Algebra. By John Stillwell. Springer-Verlag, New York, Inc., 1994, Undergraduate Texts in Mathematics, viii + 181, $34.95.

Reviewed by **Gene Freudenburg**

Historically, a large portion of what now falls under the rubric of *algebra* developed through the study of polynomial equations having rational or integral coefficients:

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = 0.$$

In the standard undergraduate mathematics curriculum, important results concerning such equations usually come as corollaries to a more general algebraic theory. From the students' point of view, this approach leaves a bewildering gap between

the abstractions of the new theory and the familiar tools of algebra used in high school to study polynomials.

Most students first encounter (non-linear) polynomial equations via an introduction to the Quadratic Formula. A course in college algebra might then present the Division Algorithm and the Factor Theorem. Finally, most texts for such a course also give a statement of the Fundamental Theorem of Algebra, possibly providing a brief sketch of its history. An ambitious text might even give the formula of Cardano for finding the roots of a third-degree polynomial, and state that no such formula can be given generally for equations of degree five or greater. (We say that an equation of the form above is "solvable by radicals" if its roots can be expressed in terms of the coefficients $a_0, \dots, a_n$ using operations $+, -, \times, \div$ and radicals $\sqrt{\cdot}, \sqrt[3]{\cdot}, \dots$ in a finite number of steps.)

Typically, a student of mathematics next encounters the study of polynomials in an undergraduate course in modern algebra or Galois theory. Such courses focus on groups and fields, objects which were created to study polynomial equations, but which have themselves become the objects of study. Indeed, these tools are indispensable in proving the unsolvability by radicals of the general n^{th} degree equation for $n \geq 5$. But there remains a large gap, both pedagogical and historical, between the elementary tools of college algebra and the abstract constructs of modern algebra used in the study of polynomial equations.

In *Elements of Algebra*: *Geometry*, *Numbers*, *Equations*, John Stillwell accomplishes, among other things, the bridging of this gap. The author strives to show why certain ideas of algebra—groups, rings, fields, and dimension—are *natural* as well as powerful. To this end, the writing is organized around the historical development of algebra, although the primary undertaking of the author is mathematics, not history. While such a work could easily become overloaded with "information", both historical and mathematical, Stillwell has taken pains to keep his presentation focused and concise.[†] Moreover, the reader is continually invited to draw on his or her own intuition to *anticipate* key developments prior to their formal exposition, to think through the same difficult questions as did Gauss, Abel, and Galois.

Another stated purpose of the book is to re-establish the connection between algebra and geometry seemingly lost in the refinement of algebraic ideas over the last century. In doing so, the author addresses a question left unanswered in many algebra texts, namely the question as to *why* understanding polynomial equations is an important mathematical endeavor. The first chapter (of nine chapters) is largely devoted to describing classical construction problems of Euclidean geometry. We read: "the most important point to observe is that diverse geometric problems reduce to the solution of polynomial equations. Understanding polynomial equations is, therefore, a more fundamental problem, and of interest in its own right". Likewise, in subsequent chapters which examine the development of number theory and number systems, the author writes, "it is probably necessary to reiterate that ring theory grew out of the attempt to model the theory of algebraic integers on the theory of ordinary integers".

[†]In his recent article *Galois Theory for Beginners* (Monthly, Vol. 101, No. 1), Stillwell gives a proof of the unsolvability of the general quintic equation, and writes: "most of the standard approach had to be stripped away before the present proof became visible. I read the books of Edwards, Tignol, Artin, Kaplansky, Maclane and Birkhoff, and Lang, taught a course in Galois theory, and then discarded 90% of what I had learned."

Mathematically, one of the main goals of the book is proving that the general n^{th} degree equation is not solvable by radicals when $n \geq 5$. Since most of the fundamental concepts of group theory are required for Galois theory, about half of the book (the last half) is devoted to developing these concepts. Clearly, any work seeking to relate ideas of modern algebra to their historical origins must come to terms with the abstract nature of the former. Stillwell concedes: "Even the present treatment is not very faithful to the history of the subject, because some modern concepts are simply too efficient to do without." Nonetheless, in his exposition the author is careful to avoid abstraction for its own sake, stating that algebra provides "the best and purest form of application because it reveals the simplest and most universal mathematical structures". Stillwell acts as a narrator who sets the stage for new definitions and theorems along the way, reminding us, for example, that the terms "solvable group" and "normal subgroup" have their origin in the attempt to solve equations by radicals. The progression of ideas is described nicely in the following passage.

> In short, a solution of the general n^{th} degree equation by radicals entails a radical extension $\overline{E}$ of the coefficient field $\mathbb{Q}(a_0, \ldots, a_{n-1})$ which is "fully symmetric" in $x_1, \ldots, x_n$. On the other hand, the radical extensions we know from Section 8.1 seem to have rather limited symmetry, with each radical extending the Galois group by only a small step. This suggests a way to prove *nonexistence* of solutions by radicals, at least for $n \geq 5$, by showing that radical extensions have less than full symmetry in $x_1, \ldots, x_n$. In Section 8.3, we shall show that the Galois group $\mathrm{Gal}(F(\alpha_1, \ldots, \alpha_k): F)$ of any radical extension has a special structure, called *solvability*, inherited from the sequence of adjoined radicals α_i. Then in Section 8.4, we shall show that this structure precludes full symmetry in $x_1, \ldots, x_n$, when $n \geq 5$, thus completing the proof.

Stillwell divides the book into short sections, with a few (at most seven) exercises at the end of each section. Each chapter is concluded with a two- to three-page "Discussion", wherein the author gives an historical overview of the main ideas of the chapter, and also discusses recent developments and open problems. These discussions contribute greatly to the overall effectiveness of the book.

As a text, the book is well-suited for a one-semester course, and should be both accessible and appealing to a wide range of students. The structure of the book is flexible enough to allow for its use in a shorter time period (e.g., an academic quarter), or with non-mathematics majors having a good understanding of college algebra and the concepts of function and set theory. It is also undoubtedly a good resource for students who have already studied Galois theory, providing historical context and clarification for the mathematical ideas.

Finally, it should be pointed out that the study of polynomials is a topic of renewed interest and importance, due in part to the advent of computer algebra systems. Using the recently-developed Method of Groebner Bases, many computational problems can be reduced to solving polynomial equations in one variable. For example, consider a system of n polynomial equations in n variables:

$$\begin{aligned} p_1(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ p_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

(Such a system often arises in the use of Lagrange multipliers in third-semester calculus.) In many cases, application of Buchberger's Algorithm transforms the

system into one of the form:

$$\begin{aligned} q_1(x_1) &= 0 \\ q_2(x_1, x_2) &= 0 \\ &\vdots \\ q_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

where $q_1, \ldots, q_n$ are again polynomials. (This important algorithm was originally developed by Bruno Buchberger in 1965, and is equivalent to Gaussian elimination when the given system is linear.) So we are again faced with the age-old question of finding the roots of $q_1(x_1)$; knowing the roots $\lambda_1, \ldots, \lambda_m$ of q_1, we substitute these for x_1 in $q_2(x_1, x_2)$, and must then find the roots of $q_2(\lambda_i, x_2)$ for each i; and so on.

In summary, *Elements of Algebra* is an engaging and inspiring account of the main ideas of algebra in which Stillwell successfully conveys his own insight, knowledge, and enthusiasm for algebra and its history. The book is intended as a self-contained introduction to abstract algebra for undergraduates, one which emphasizes the role of algebra in geometry and number theory in order to motivate the algebraic theory. The author also points the way for further pursuit of the topics discussed, since "the aim of the book is to lead the reader to better things".

Department of Mathematics
University of Southern Indiana
Evansville, IN 47712

At the age of eleven, I began Euclid, with my brother as my tutor. This was one of the great events of my life, as dazzling as first love. I had not imagined there was anything so delicious in the world... From that moment until... I was thirty-eight, mathematics was my chief interest and my chief source of happiness.

—*Bertrand Russell (1872–1970)*

The Autobiography of Bertrand Russell,
London: G. Allen and Unwin, 1968.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee and Paul Zorn**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S . Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *Symposia Gaussiana, Conference A: Mathematics and Theoretical Physics.* Eds: Minaketan Behara, Rudolf Fritsch, Rubens G. Lintz. Walter de Gruyter, 1995, xx + 744 pp, DM 328. [ISBN 3-11-014476-X] Proceedings of a 1993 symposium in Munich, Germany. Topics include mathematics education, history of mathematics, mathematical logic, number theory, geometry, analysis, algebra, and mathematical physics.

Mathematics Appreciation, L.** *In Search of Infinity.* N. Ya. Vilenkin. Transl: Abe Shenitzer. Birkhäuser Boston, 1995, vii + 145 pp, $24.50. [ISBN 0-8176-3819-9] A wide-ranging history of set theory with engaging examples and lucid explanations; written for a general audience. Begins with early concepts of infinity, Zeno's paradoxes, origins of the concept of function. Describes infinite cardinals, non-Euclidean geometries, pathological functions, space-filling curves, Gödel's incompleteness theorem, and much more. Highlights contributions of Russian mathematicians. DB

Mathematics Appreciation, T(13: 1). *Math Matters.* James V. Rauff. Wiley, 1996, xv + 491 pp, $40.95 (P). [ISBN 0-471-30452-2] In workbook format, introduces mathematics in "real life" (e.g., ratios and proportions appear in context of price markdowns). Includes very basic introductions to probability and statistics, linear programming, matrices, logic and "mathematics of computers." Assumes only one year of high school algebra. JNC

Recreational Mathematics, S*(13–16), L*. *Colorado Mathematical Olympiad: The First 10 Years and Further Explorations.* Alexander Soifer. Center for Excellence in Math Educ (885 Red Mesa Dr, Colorado Springs, CO 80906), 1994, x + 189 pp, $19.95 (P). [ISBN 0-940263-03-3] A compendium of accessible problems, solutions, and statistics on the results. Distinguished by personal and conversational style; with historical notes and anecdotes, related conjectures, and open problems. LCL

Elementary, T(13–14: 1). *Modern Mathematics for Elementary School Teachers, Ninth Edition.* Ruric E. Wheeler, Ed R. Wheeler. Brooks/Cole, 1995, xvi + 919 pp. [ISBN 0-534-25326-1] New features, added in response to NCTM *Standards*: (1) "PCR Excursions" encourage active investigation, problem solving, communication, and reasoning; (2) sections begin with statements from the *Standards*; (3) exercises and activities use graphing calculators and dynamic drawing software. MW

Education, P. *Teachers' Minds and Actions: Research on Teachers' Thinking and Practice.* Eds: Ingrid Carlgren, Gunnar Handal, Sveinung Vaage. Falmer Pr, 1994, viii + 279 pp, $26 (P). [ISBN 0-7407-0431-4] 17 papers from a 1994 conference in Gothenburg, Sweden, sponsored by the International Study Association on Teacher Thinking.

Education, P. *Beliefs and Values in Science Education.* Michael Poole. Open Univ Pr, 1995, 146 pp, $24.95 (P). [ISBN 0-335-15646-0] A collage of disconnected vignettes and episodes from the interface of science and society, confronting directly in a British context what U.S. science educators generally avoid:

direct dialogue about the relation of belief systems inherent in science to the religious and ethical beliefs of ordinary citizens. LAS

Education, P. *Changing the Culture: Mathematics Education in the Research Community.* Eds: Naomi D. Fisher, Harvey B. Keynes, Philip D. Wagreich. CBMS Issues in Math. Educ., V. 5. AMS and MAA, 1995, x + 213 pp, $59 (P). [ISBN 0-8218-0383-2] 14 essays by participants in Mathematicians and Education Reform (MER) Network programs.

Education, S, P. *Ethnomathematics: A Multicultural View of Mathematical Ideas.* Marcia Ascher. Chapman & Hall, ix + 203 pp, (P). [ISBN 0-412-98941-7] Paperback republication of 1991 Brooks/Cole edition (TR, August–September 1991; Extended Review, March 1993).

History, P, L. *The Scientific Letters and Papers of James Clerk Maxwell, Volume II, 1862–1873.* Ed: P.M. Harman. Cambridge Univ Pr, 1995, xxx + 999 pp, $285. [ISBN 0-521-25626-7]

History, S, L. *Classics of Mathematics.* Ed: Ronald Calinger. Prentice Hall, 1995, xxi + 793 pp, (P). [ISBN 0-02-318342-X] Anthology includes selections by Proclus, Euclid, Plato, Oresme, Pascal, Fermat, Newton, Euler, Hilbert, and many others. Historical and biographical notes precede each chapter. Valuable historical resource. (1982 Moore edition, TR, January 1984; Extended Review, October 1985.) LC

History, S, L. *Out of Their Minds: The Lives and Discoveries of 15 Great Computer Scientists.* Dennis Shasha, Cathy Lazere. Springer-Verlag, 1995, xi + 291 pp, $23. [ISBN 0-387-97992-1] Brief, interview-based biographies of 15 computer scientists. Includes glossary, bibliography, interviewees' favorite publications, a small computer science timeline. JO

History, S(15–16), P*, L.** *Force and Geometry in Newton's Principia.* François De Gandt. Princeton Univ Pr, 1995, xiv + 296 pp, $49.50. [ISBN 0-691-03367-6] An essential companion for exploring the physics or mathematics of Newton's *Principia.* Clear, careful treatment of Newton's concept of force and its historical antecedents. Illuminating explanation of Newton's use of infinitesimal analysis and its historical context. Provides rich insights into the origins of calculus. DB

Logic, P. *Advances in Linear Logic.* Eds: Jean-Yves Girard, Yves Lafont, Laurent Regnier. London Math. Soc. Lect. Note Ser., V. 222. Cambridge Univ Pr, 1995, vii + 389 pp, $44.95 (P). [ISBN 0-521-55961-8] Proceedings of a 1993 Cornell University conference.

Logic, P. *Set Theory: On the Structure of the Real Line.* Tomek Bartoszyński, Haim Judah. AK Peters, 1995, xi + 546 pp, $69.96. [ISBN 1-56881-044-x] Presents recent results in descriptive set theory that can be proven using the Zermelo-Fraenkel axioms. DB

Combinatorics, P. *Davenport-Schinzel Sequences and Their Geometric Applications.* Micha Sharir, Pankaj K. Agarwal. Cambridge Univ Pr, 1995, xii + 372 pp, $49.95. [ISBN 0-521-47025-0] Davenport-Schinzel sequences arise in the study of envelopes of collections of functions, and are useful in discrete and computational geometry. LC

Number Theory, P. *Number Theory: Séminaire de Théorie des Nombres de Paris 1992–3.* Ed: Sinnou David. London Math. Soc. Lect. Note Ser., V. 215. Cambridge Univ Pr, 1995, 291 pp, $39.95 (P). [ISBN 0-521-55911-1]

Number Theory, T(18), P*. *Modular Forms and Hecke Operators.* A.N. Andrianov, V.G. Zhuravlev. Transl. of Math. Mono., V. 145. AMS, 1995, vii + 334 pp, $95. [ISBN 0-8218-0277-1] Given a quadratic form in an even number of variables, the number of ways it can represent a given integer is a linear combination of multiplicative functions. The formulas are often striking. Why these formulas arise and how to find them was explained by Hecke in 1937. To accomplish this, he introduced operators for studying multiplicative properties of the Fourier coefficients of modular forms. This book, describing Hecke's theory and associated work, can serve as an introduction to theta series and modular forms. DB

Linear Algebra, T(14). *Elementary Linear Algebra with Applications, Third Edition.* Richard O. Hill, Jr. Saunders College, 1996, xvii + 516 pp, $51. [ISBN 0-03-010347-9] Standard linear algebra text with interesting applications. An option permits reducing the role of determinants. (*Second Edition*, TR, November 1991.) PF

Group Theory, P. *Conjugacy Classes in Semisimple Algebraic Groups.* James E. Humphreys. Math. Surveys & Mono., V. 43. AMS, 1995, xviii + 196 pp, $59. [ISBN 0-8218-0333-6]

Group Theory, P. *Geometric Group Theory.* Eds: Ruth Charney, Michael Davis, Michael Shapiro. Walter de Gruyter, 1995, ix + 186 pp, DM 148. [ISBN 3-11-014743-2] Proceed-

ings of a special research quarter at The Ohio State University in spring 1992.

Algebra, P. *Lie Groups and Lie Algebras: E.B. Dynkin's Seminar.* Eds: S.G. Gindikin, E.B. Vinberg. Transl. Ser. 2, V. 169: Adv. in Math. Sci., 26. AMS, 1995, xi + 202 pp. [ISBN 0-8218-0454-5]

Algebra, P. *Computational Algebra and Number Theory.* Eds: Wieb Bosma, Alf van der Poorten. Math. & Its Applic., V. 325. Kluwer Academic, 1995, xiv + 321 pp, $156. [ISBN 0-7923-3501-5] 22 papers from a 1992 conference at Sydney University.

Algebra, P. *Representation Theory and Harmonic Analysis.* Eds: Tuong Ton-That, *et al.* Contemp. Math., V. 191. AMS, 1995, xii + 254 pp, $55 (P). [ISBN 0-8218-0310-7] Proceedings of a Special Session at the January 1994 AMS meeting in Cincinnati.

Algebra, P. *Recent Developments in the Inverse Galois Problem.* Eds: Michael D. Fried, *et al.* Contemp. Math., V. 186. AMS, 1995, x + 401 pp, $65 (P). [ISBN 0-8218-0299-2] Proceedings of a 1993 conference at the University of Washington.

Algebra, P. *Semigroups on Algebra, Geometry and Analysis.* Eds: Karl H. Hofmann, Jimmie D. Lawson, Ernest B. Vinberg. Expos. in Math., V. 20. Walter de Gruyter, 1995, xii + 370 pp, DM 198. [ISBN 3-11-014319-4] 14 survey papers, in 6 sections: Lie semigroups, ordered symmetric spaces, and causality; invariant cones, Ol'shankskiĭ semigroups, exponential subgroups; convexity theorems and representation theory; semisimple Lie groups and semigroups; applications to control; applications to probability.

Calculus, T(14). *Multivariable Calculus, Fifth Edition.* Howard Anton. Wiley, 1995, xxvii + 446 pp, $62.95 (P). [ISBN 0-471-13909-2] Multivariable portion of *Calculus with Analytic Geometry, 5th Edition* (TR, October 1995). Also material on differential equations, infinite series, Cramer's rule, and complex numbers.

Real Analysis, T(15). *An Introduction to Analysis, Second Edition.* James R. Kirkwood. PWS, 1995, vii + 278 pp. [ISBN 0-534-94422-1] Like the *First Edition* (TR, October 1989), this is a clean, readable little text that covers the first topics to be taken up after calculus. In fact, this is the *First Edition*, all over again, called a *Second Edition* for barely perceptible reasons. AWR

Complex Analysis, T(17–18), P. *Elementary Theory of Analytic Functions of One or Several Complex Variables.* Henri Cartan. Dover, 1995, 227 pp, $8.95 (P). [ISBN 0-486-68543-8] Unabridged republication of the 1973 Addison-Wesley edition.

Partial Differential Equations, P. *The Interplay between Differential Geometry and Differential Equations.* Ed: V.V. Lychagin. Transl. Ser. 2, V. 167. AMS, 1995, ix + 294 pp. [ISBN 0-8218-0428-6]

Partial Differential Equations, P. *KdV '95.* Eds: Michiel Hazewinkel, Hans W. Capel, Eduard M. de Jager. Kluwer Academic, 1995, vi + 516 pp, $269. [ISBN 0-7923-3467-1] Proceedings of a 1995 symposium in Amsterdam commemorating the centennial of the Korteweg-de Vries equation.

Partial Differential Equations, T(18: 2), P. *Boundary Value Problems for Elliptic Systems.* J.T. Wloka, B. Rowley, B. Lawruk. Cambridge Univ Pr, 1995, xiv + 641 pp, $89.95. [ISBN 0-521-43011-9] Graduate-level text. "The aim is to simplify and to algebraize the index theory by means of pseudo-differential operators and new methods in the spectral theory of matrix polynomials." LC

Dynamical Systems, P. *Sinai's Moscow Seminar on Dynamical Systems.* Eds: L.A. Bunimovich, B.M. Gurevich, Ya. B. Pesin. Transl. Ser. 2, V. 171: Adv. in Math. Sci., 28. AMS, 1996, xi + 247 pp, $95. [ISBN 0-8218-0426-1]

Dynamical Systems, T, S(18), P*. *Applied Analysis of the Navier-Stokes Equations.* Charles R. Doering, J.D. Gibbon. Texts in Appl. Math. Cambridge Univ Pr, 1995, xiii + 217 pp, $24.95 (P); $59.95. [ISBN 0-521-44568-X; 0-521-44557-4] The Navier-Stokes equations are the standard PDE models for fluid motion of gases and liquids from laminar to turbulent flow. Their solutions are still poorly understood. It is not even known whether smooth or even unique solutions must follow from smooth initial conditions. This book explains what *is* known. Addresses a general audience of mathematicians, physicists, and engineers with the equivalent of one year of graduate mathematics. Includes exercises. DB

Dynamical Systems, P. *Real and Complex Dynamical Systems.* Eds: Bodil Branner, Poul Hjorth. NATO ASI Ser. C, V. 464. Kluwer Academic, 1995, xvii + 344 pp, $174. [ISBN 0-7923-3521-X] Proceedings of 1993 NATO Advanced Study Institute in Denmark.

Numerical Analysis, P. *Acta Numerica 1995.* Ed: A. Iserles. Cambridge Univ Pr, 1995, 491 pp, $59.95. [ISBN 0-521-48255-0] Survey papers on quadratic programming, Padé ap-

proximation, differential equations, and computation of eigenvalues and eigenvectors.

Functional Analysis, T(17–18), S. *Applied Functional Analysis: Main Principles and Their Applications.* Eberhard Zeidler. Appl. Math. Sci., V. 109. Springer-Verlag, 1995, xvi + 404 pp, $59. [ISBN 0-387-94422-2] Second book in a two-part introduction to applied functional analysis, but basic knowledge of Banach spaces suffices (*Applications to Mathematical Physics*, AMS V. 108, is first part). Motivates many standard topics (Hahn-Banach theorem, open mapping theorem, implicit function theorem, Fredholm operators) with impressive array of applications. Stresses problem solving more than mathematical completeness. SA

Functional Analysis, T(17–18), S, P. *Finite Sums Decompositions in Mathematical Analysis.* Themistocles M. Rassias, Jaromír Šimša. Wiley, 1995, vi + 172 pp, $49.95. [ISBN 0-471-94827-6] Summary of classical and current results relating to the problem of decomposing a function of several variables into a finite sum of products of functions of a single variable. Includes applications to PDE's, integral equations, and L^2-approximation theory. Assumes basic knowledge of functional analysis, differential geometry. SA

Functional Analysis, T(18: 2), P. *Spectral Methods in Infinite-Dimensional Analysis.* Y.M. Berezansky, Y.G. Kondratiev. Kluwer Academic, 1995, $495 set. *Volume 1*, Math. Physics & Appl. Math., V. 12/1, xvii + 572 pp, [ISBN 0-7923-2847-7]; *Volume 2*, Math. Physics & Appl. Math., V. 12/2, viii + 427 pp. [ISBN 0-7923-2848-5] English version of the Russian original. Centers on study of generalized functions of infinitely many variables. *Volume 1* (Chapters 1–4) develops the spectral theory needed for the applications in *Volume 2* (Chapters 5–7). Applications include infinite-dimensional moment problems, positive definite functions, elliptic differential operators, quantum statistical physics, and field theory. Bibliography, notes, index. JS

Functional Analysis, P. *Classification of Subfactors and Their Endomorphisms.* Sorin Popa. CBMS, No. 86. AMS, 1995, x + 110 pp, (P). [ISBN 0-8218-0321-2]

Analysis, P. *Harmonic Analysis in China.* Eds: Minde Cheng, *et al.* Math. & Its Applic., V. 327. Kluwer Academic, 1995, x + 307 pp, $163. [ISBN 0-7923-3566-X] 17 papers surveying recent research in China.

Analysis, T(15–16). *The Way of Analysis.* Robert S. Strichartz. Jones & Bartlett, 1995, xix + 739 pp, $62.50. [ISBN 0-86720-471-0] Undergraduate real analysis, supplemented with material on metric spaces, complex analysis, differential equations, vector analysis, and measure theory. DB

Analysis, P. *Harmonic Analysis and Operator Theory.* S.A.M. Marcantognini, *et al.* Contemp. Math., V. 189. AMS, 1995, xi + 511 pp, $75 (P). [ISBN 0-8218-0304-2] Proceedings of a 1994 conference in Caracas, Venezuela, honoring Mischa Cotlar.

Analysis, T(17–18), S, P. *Navier-Stokes Equations in Irregular Domains.* L. Stupelis. Math. & Its Applic., V. 326. Kluwer Academic, 1995, xv + 566 pp, $249. [ISBN 0-7923-3509-0] Extensive treatment of boundary value problems for Stokes and Navier-Stokes equations in domains with smooth and irregular boundaries. Includes introductions to Banach spaces and function spaces, e.g., Sobolev and Hölder spaces. Applications to motion of a viscous incompressible fluid in an open container. SA

Analysis, P. *Fourier Transforms.* Ian N. Sneddon. Dover, 1995, xii + 542 pp, $14.95 (P). [ISBN 0-486-68522-5] Unabridged republication of the 1951 McGraw-Hill edition.

Algebraic Geometry, P. *Arithmetic of Diagonal Hypersurfaces over Finite Fields.* Fernando Q. Gouvêa, Noriko Yui. London Math. Soc. Lect. Note Ser., V. 209. Cambridge Univ Pr, 1995, xi + 169 pp, $32.95 (P). [ISBN 0-521-49834-1]

Algebraic Geometry, T(18: 1), P. *Shafarevich Maps and Automorphic Forms.* János Kollár. Princeton Univ Pr, 1995, ix + 201 pp, $37.50. [ISBN 0-691-04381-7] Elaborates connections between the theory of automorphic forms and a weakened form of the Shafarevich conjecture (that the universal cover of smooth projective algebraic variety is holomorphically convex). Aims at study of varieties with infinite fundamental groups. RM

Algebraic Geometry, P. *Automorphisms of Affine Spaces.* Ed: Arno van den Essen. Kluwer Academic, 1995, xix + 243 pp, $132. [ISBN 0-7923-3523-6] Proceedings of a 1994 Caribbean Mathematical Foundation conference in Curaçao.

Geometry, T(18: 1), P. *Equivalence, Invariants, and Symmetry.* Peter J. Olver. Cambridge Univ Pr, 1995, xvi + 525 pp, $39.95. [ISBN 0-521-47811-1] A clearly written *tour de force.* Systematically explores problems of equivalence (determining when two objects are identical under change of variables), symmetry ("self-equivalence"), and classification of in-

variants of geometric objects. Treats problems in geometry, differential equations, and the calculus of variations, using analytic, continuous, and local methods. Makes nice connections among many topics. RM

Algebraic Topology, P. *Homotopy Theory and Its Applications.* Eds: Alejandro Adem, R. James Miligram, Douglas C. Ravenel. Contemp. Math., V. 188. AMS, viii + 229 pp, $49 (P). [ISBN 0-8218-0305-0] Proceedings of a 1993 conference in Cocoyoc, Mexico, honoring Samuel Gitler.

Topology, T(15–17), P. *Topology: An Introduction to the Point-Set and Algebraic Areas.* Donald W. Kahn. Dover, 1995, viii + 217 pp, $7.95 (P). [ISBN 0-486-68609-4] Unabridged, corrected republication of the 1975 Williams & Wilkins edition (TR, December 1975).

Optimization, P. *Recent Developments in Well-Posed Variational Problems.* Eds: Roberto Lucchetti, Julian Revalski. Math. & Its Applic., V. 331. Kluwer Academic, 1995, viii + 266 pp, $145. [ISBN 0-7923-3576-7] Articles on approximate solutions, well-posedness and stability of problems in scalar and vector optimization, game theory, and calculus of variations.

Elementary Statistics, T(13: 1), S. *Statistics for the Terrified.* Gerald Kranzler, Janet Moursund. Prentice Hall, 1995, x + 164 pp, (P). [ISBN 0-13-183831-8] Aimed at students with weak algebra skills. A brief introduction, reducing statistical tests to step-by-step formulas. Brief descriptions tell when to use each test. Many topics briefly treated: data summaries, correlation, inferential statistics, regression, difference between two means, one-way ANOVA, non-parametric tests. Includes solved problems, statistics tables, basic math review, advice for the math-anxious. KB

Statistical Methods, P. *Semiparametric Models in Accelerated Life Testing.* V. Bagdonavičius, M. Nikulin. Papers in Pure & Appl. Math., No. 98. Queen's Univ, 1995, ii + 70 pp, (P). [ISBN 0-88911-726-8]

Statistical Methods, T(17–18: 2), P, L. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments.* Raymond H. Myers, Douglas C. Montgomery. Ser. in Prob. & Stat. Wiley, 1995, xiv + 700 pp, $59.95. [ISBN 0-471-58100-3] An RSM provides an empirical model for a response based on predictor variables when the exact relationship is unknown. Assumes knowledge of experimental design, regression modeling, elementary optimization methods, matrix algebra. Stresses methods useful in industry. Includes examples of computer data analysis and experimental design. Chapter problems (no answers). KB

Statistical Methods, S(18), P*. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* Robert E. Bechhofer, Thomas J. Santner, David M. Goldsman. Ser. in Prob. & Stat. Wiley, 1995, xii + 325 pp, $54.95. [ISBN 0-471-57427-9] Assumes knowledge of standard experimental design. Discusses selection procedures using the indifference-zone approach and screening procedures using the subset approach. Selection procedures are applied to normal, Bernoulli, and multinomial probability models. Develops multiple comparison procedures using normal means. Provides tables and Fortran programs to implement procedures. Readable, with many examples but no exercises. RS

Statistical Methods, T(18: 1), P. *Observational Studies.* Paul R. Rosenbaum. Ser. in Stat. Springer-Verlag, 1995, xv + 230 pp, $44.95. [ISBN 0-387-94482-6] Observational studies assess effects of treatments when controlled experiments are not feasible. Derives optimal matching and stratification strategies. Treats bias detection, sensitivity analysis, selection bias, poset statistics for tests of coherent association, and propensity scores in permutation inference. Some sections are accessible to undergraduates. RS

Statistics, P. *Symposia Gaussiana, Conference B: Statistical Sciences.* Eds: Volker Mammitzsch, Hans Schneeweiß. Walter de Gruyter, 1995, x + 341 pp, DM 268. [ISBN 3-11-014412-3] Proceedings of a 1993 symposium in Munich. Papers on probability theory, expert systems, decision theory, simulation, and design of experiments.

Statistics, T(18: 2), P, L. *Asymptotic Efficiency of Nonparametric Tests.* Yakov Nikitin. Cambridge Univ Pr, 1995, xvi + 274 pp, $49.95. [ISBN 0-521-47029-3] Monograph on analysis and calculation of asymptotic efficiencies of nonparametric tests. Methods are based on Sanov's theorem, variational calculus, and nonlinear analysis. Finds the Bahadur, Hodges-Lehmann, and Chernoff efficiencies for many nonparametric tests for goodness-of-fit, homogeneity, symmetry, and independence hypotheses. No exercises. KB

Computer Systems, C, P. *Netscape Navigator: Surfing the Web and Exploring the Internet.* Bryan Pfaffenberger. Academic Pr, 1995. *Window's Version*, xx + 362 pp, $29.95 (P), with CD-ROM, [ISBN 0-12-553132-X]; *Mac-*

intosh Version, xx + 314 pp, \$29.95 (P), with CD ROM. [ISBN 0-12-553131-1]

Computer Systems, P. *Practical C++ Programming.* Steve Oualline. O'Reilly & Assoc, 1995, xxiv + 557 pp, \$24.95 (P). [ISBN 1-56592-139-9]

Computer Systems, C, P. *C++ Programming with CodeWarrior: Beginning OOP for the Macintosh and Power Macintosh.* Jan L. Harrington. Academic Pr, 1995, xxiii + 373 pp, \$34.95 (P), with CD ROM. [ISBN 0-12-326420-0]

Computer Systems, C, P. *Agents Unleashed: A Public Domain Look at Agent Technology.* Peter Wayner. Academic Pr, 1995, xii + 358 pp, \$39.95 (P), with disk. [ISBN 0-12-738765-X]

Computer Systems, P. *Microsoft RPC Programming Guide.* John Shirley, Ward Rosenberry. O'Reilly & Assoc, 1995, xix + 232 pp, \$24.95 (P). [ISBN 1-56592-070-8]

Computer Science, T(13: 1). *Problem Solving and Programming Concepts, Third Edition.* Maureen Sprankle. Prentice Hall, 1995, xv + 471 pp, (P). [ISBN 0-02-415350-8] Intended for "language or application majors." Essentially an introductory programming course without the programs. Examples and exercises use flowcharts and miscellaneous planning diagrams. Includes chapters on problem solving with word processors, spreadsheets, and database managers. JO

Computer Science, P. *The NURBS Book.* Les Piegl, Wayne Tiller. Mono. in Visual Communication. Springer-Verlag, 1995, xiv + 646 pp, \$89. [ISBN 3-540-55069-0] Comprehensive reference on non-uniform rational B-splines. Geometric and numerical-analytic background, thousands of lines of nearly-C code, and lots of programming advice. If you want to know about NURBS, start here. JO

Computer Science, P. *Algorithmic Foundations of Robotics.* Eds: Ken Goldberg, *et al.* AK Peters, 1995, xii + 556 pp, \$52. [ISBN 1-56881-045-8] Proceedings of a 1994 workshop in San Francisco.

Computer Science, T(13; 1), L. *The Most Complex Machine: A Survey of Computers and Computing.* David J. Eck. AK Peters, 1995, xii + 445 pp, \$49.95. [ISBN 1-56881-054-7] An up-to-date, conceptual introduction to ideas and principles of computing. Nicely balances historical development, algorithms and languages, modern applications. Appropriate for non-majors, but worth a look by anyone serious about CS. Labs and other materials are available on the Web. RM

Applications (Biological Science), P. *Dynamical Disease: Mathematical Analysis of Human Illness.* Eds: Jacques Bélair, *et al.* American Institute of Physics, 1995, ix + 220 pp, \$40 (P). [ISBN 1-56396-370-1] 30 papers from a 1994 NATO Advanced Research Workshop in Québec.

Applications (Fluid Mechanics), S(18), P. *Navier-Stokes Equations and Nonlinear Functional Analysis, Second Edition.* Roger Temam. SIAM, 1995, xiv + 141 pp, \$26.50 (P). [ISBN 0-89871-340-4] This edition includes a new appendix on inertial manifolds and an updated bibliography. HD

Applications (Physics), P. *Affine Lie Algebras and Quantum Groups: An Introduction, with Applications in Conformal Field Theory.* Jürgen Fuchs. Mono. on Math. Physics. Cambridge Univ Pr, 1995, xiv + 433 pp, \$39.95 (P); \$100. [ISBN 0-521-48412-X; 0-521-41593-4] Paperback edition, with corrections (TR, October 1993).

Applications (Physics), P. *Topics in Topology and Mathematical Physics.* Ed: S.P. Novikov. Transl. Ser. 2, V. 170: Adv. in Math. Sci., 27. AMS, 1995, vii + 206 pp. [ISBN 0-8218-0455-3] 8 papers based on talks presented at S.P. Novikov's seminar in Moscow.

Applications (Physics), P. *Diffraction by Wedges.* Bair Budaev. Pitman Res. Notes in Math. Ser., V. 322. Longman Scientific & Technical (Copub: Wiley), 1995, 141 pp, \$49.95 (P). [ISBN 0-582-10310-X]

Applications (Physics), P. *Geometric Scattering Theory.* Richard B. Melrose. Cambridge Univ Pr, 1995, xi + 116 pp, \$14.95 (P); \$39.95. [ISBN 0-521-49810-4; 0-521-49673-X] A "non-technical" overview of scattering theory. Based on lectures given in 1994 at Stanford and MIT.

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–205: Inversion Method in the Discrete-time Nonlinear Control Systems Synthesis Problems.* Ülle Kotta. Springer-Verlag, 1995, xi + 152 pp, \$45 (P). [ISBN 3-540-19966-7]

Reviewers

SA: Steve Abbott, St. Olaf; KB: Karla Ballman, Macalester; DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; LC: Laura Chihara, St. Olaf; HD: Hung Dinh, Macalester; PF: Paul Froeschl, Macalester; LCL: Loren C. Larson, St. Olaf; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; AWR: A. Wayne Roberts, Macalester; JS: John Schue, Macalester; RS: Richard Single, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; MW: Martha Wallace, St. Olaf.

All the Math That's Fit to Print

Articles from the Manchester Guardian

Keith Devlin

Mathematics and mathematicians can be the objects of public interest, if there are individuals capable of explaining those items in a form that the intelligent reader can follow. Keith Devlin is such a person and the editors of the British paper, The Manchester Guardian, were intelligent enough to understand that. This book should be an element of every public library.

—Journal of Recreational Mathematics

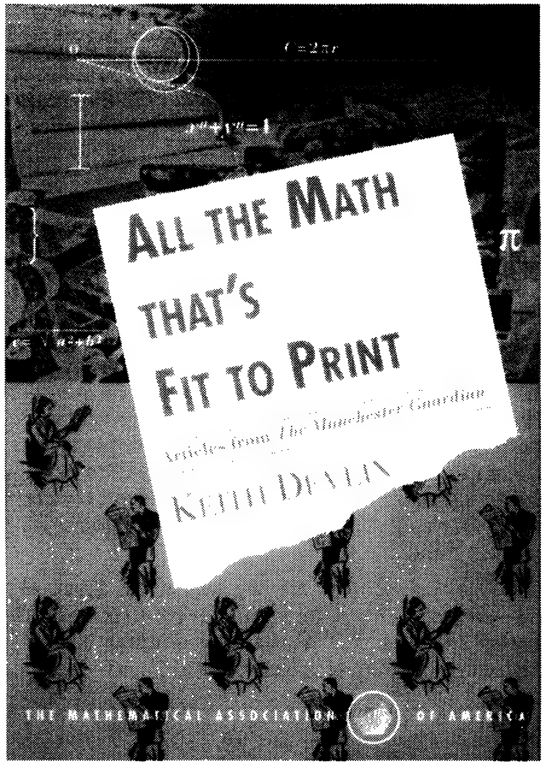
This new work reveals another side of Devlin's interesting investigations into mathematics and his efforts to share them with laypersons...Anyone interested in mathematics will find something of interest in this book...When possible, the author provides a historical context for the new ideas being explored.

—Choice

Between 1983 and 1989 Keith Devlin, research mathematician, author and educator, wrote a semi-monthly column on mathematics and computing in the English national daily newspaper, The Manchester Guardian. This book is a compilation of many of those articles. It is a witty, entertaining, easy-to-read piece of work.

The mathematical topics range from simple puzzles to deep results including open problems such as Faltings Theorem and the Riemann Conjecture. You will find articles on prime numbers, how to work out claims for traveling expenses, calculating pi, computer simulation, patterns and palindromes, cryptology, and much more.

This book is meant for browsing by anyone who regularly reads a serious newspaper and has some interest in matters scientific or mathematical.



345 pp., Paperbound, 1994
ISBN - 0-88385-515-1
List: $38.50
MAA Member: $28.95
Catalog Code: ATMA/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	ATMA/JR	______	______
______	______	______	______
		TOTAL	______

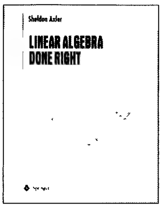
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

SPRINGER FOR MATHEMATICS

SHELDON AXLER, Michigan State University, East Lansing, MI



LINEAR ALGEBRA DONE RIGHT

This text for a second course in linear algebra is aimed at math majors and graduate students. The approach is novel, banishing determinants to the end of the book and focusing on the central goal of linear algebra: understanding the structure of linear operators on vector spaces. The author has taken unusual care to motivate concepts and to simplify proofs. For example, the book presents –without having defined determinants–a clean proof that every linear operator on a finite-dimensional complex vector space (or on an odd-dimensional real vector space) has an eigenvalue. Though this text is intended for a second course in linear algebra–following one that focuses on matrices and computations–there are no prerequisites other than appropriate mathematical maturity. Thus the book starts by discussing vector spaces, linear independence, span, basis, and dimension. Students are introduced to inner-product spaces in the first half of the book and shortly thereafter to the finite-dimensional spectral theorem. A variety of interesting exercises in each chapter helps students understand and manipulate the objects of linear algebra.

1996/238 PP./SOFTCOVER/$29.00/ISBN 0-387-94596-2
ALSO IN HARDCOVER $49.00/ISBN 0-387-94595-4
UNDERGRADUATE TEXTS IN MATHEMATICS

JOHN L. TROUTMAN, Syracuse University, Syracuse, NY

VARIATIONAL CALCULUS AND OPTIMAL CONTROL

Optimization with Elementary Convexity

This book supplies a broad-based introduction to variational methods for formulating and solving problems in mathematics and the applied sciences. It is the only current introductory text that uses elementary partial convexity of differentiable functions to characterize directly the solutions of minimization problems before exploring necessary conditions for optimality or field theory methods of sufficiency. Through effective notation, it combines rudiments of analysis in (normed) linear spaces with simpler aspects of convexity, offering a multi-level strategy for handling such problems. It also employs convexity considerations to broaden the discussion of Hamilton's principle in mechanics and to introduce Pontjragin's principle in optimal control. It is mathematically self-contained and uses applications from many disciplines to provide a wealth of examples and exercises. The book is accessible to upper-level undergraduates and will lay the groundwork for further explorations in mathematics, physics, mechanical and electrical engineering, and computer science.

1996/460 PP., 87 ILLUS./HARDCOVER/$49.95/ISBN 0-387-94511-3
UNDERGRADUATE TEXTS IN MATHEMATICS

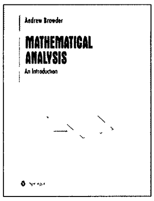
GEORGE R. EXNER, Bucknell University, Lewisburg, PA

AN ACCOMPANIMENT TO HIGHER MATHEMATICS

This text prepares undergraduate mathematics students to meet two challenges in the study of mathematics, *viz.*, to read mathematics independently and to understand and write proofs. The book begins by teaching how to read mathematics actively, constructing examples, extreme cases, and non-examples to aid in understanding an unfamiliar theorem or definition (a technique familiar to any mathematician, but rarely taught); it provides practice by indicating explicitly where work with pencil and paper must interrupt reading. The book then turns to proofs, showing in detail how to discover the structure of a potential proof from the form of the theorem (especially the conclusion). It shows the logical structure behind proof forms (especially quantifier arguments), and analyzes thoroughly, the often sketchy coding of these forms in proofs as they are ordinarily written. The common introductory material (such as sets and functions) is used for the numerous exercises, and the book concludes with a set of "Laboratories" in which the student can practice the skills learned in the earlier chapters on these topics. Intended for use as a supplementary text in courses on introductory real analysis, advanced calculus, abstract algebra, or topology, the book may also be used as the main text for a "transitions" course bridging the gap between calculus and higher mathematics.

1996/APP. 224 PP./HARDCOVER/$29.95/ISBN 0-387-94617-9
UNDERGRADUATE TEXTS IN MATHEMATICS

ANDREW BROWDER, Brown University, Providence, RI



MATHEMATICAL ANALYSIS

An Introduction

This is a textbook containing more than enough material for a year-long course in analysis at the advanced undergraduate or beginning graduate level. It begins with a brief discussion of sets and mappings, describes the real number field, and proceeds to a treatment of real-valued functions of a real variable; chapters are devoted to the ideas of convergent sequences and series, continuous functions, differentiation, and the Riemann integral. A chapter on general topology follows, as well as a chapter containing a miscellany of applications, including the Weierstrass and Stone-Weierstrass approximation theorems, the existence of geodesics in compact metric spaces, elements of Fourier analysis, and the Weyl equidistribution theorem. Subsequently, there is a discussion of differentiation of vector-valued functions of several real variables, followed by a brief treatment of measure and integration (in a general setting, but with emphasis on Lebesgue theory in Euclidean space). The final part of the book deals with manifolds, differential forms, and Stokes' theorem, which is applied to prove Brouwer's fixed point theorem and to derive the basic properties of harmonic functions, such as the Dirichlet principle.

1996/333 PP., 10 ILLUS./HARDCOVER/$39.00/ISBN 0-387-94614-4
UNDERGRADUATE TEXTS IN MATHEMATICS

VISIT THE SPRINGER WEBSITE!

Springer America {http://www.springer-ny.com}
Springer Europe {http://www.springer.de}

ORDER TODAY!

Call: 1-800-SPRINGE(R); In NJ 201-348-4033 • **Write:** Send payment (check or credit card) plus $3.00 postage and handling for the first book and $1.00 for each additional book to: Springer-Verlag New York, Inc., Dept. #S219, PO Box 2485, Secaucus, NJ 07096-2485 (CA, IL, MA, NJ, NY, PA, TX, VA, and VT residents add sales tax, Canadian resi-dents add 7% GST) • **Visit:** Your local technical bookstore • **Instructors:** Call or write for information on textbook examination copies.

2/96 Reference. S219

Springer

THE AMERICAN MATHEMATICAL MONTHLY

Volume 103, Number 3 March 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

Richard BUMBY,
P.O. Box 10971
New Brunswick, NJ 08906-0971.

Please send 2 copies of all material, typewritten if possible.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Sums of Powers of Integers

A. F. Beardon

1. INTRODUCTION. Our starting point is the well-known identity

$$1^3 + 2^3 + \cdots + n^3 = (1 + 2 + \cdots + n)^2. \tag{1.1}$$

Sums of the form

$$\sigma_k(n) = 1^k + 2^k + \cdots + n^k$$

have been studied for hundreds of years and even now there is still a steady stream of notes published on the subject, many of which can be found by browsing through back issues of the *Monthly* and the *Mathematical Gazette*. Most of these articles are concerned with generalizing (1.1), now written as

$$\sigma_3 = \sigma_1^2,$$

either by expressing a power of σ_1 as a linear combination of powers of other σ_i, for example,

$$\sigma_1^3 = \frac{1}{4}\sigma_3 + \frac{3}{4}\sigma_5, \qquad \sigma_1^5 = \frac{1}{16}\sigma_5 + \frac{5}{8}\sigma_7 + \frac{5}{16}\sigma_9, \tag{1.2}$$

or with identities involving σ_k and binomial coefficients, for example,

$$\sigma_2(n) = 2\binom{n+1}{3} + \binom{n+1}{2}, \quad \sigma_5(n) = \binom{n+1}{2} + 30\binom{n+2}{4} + 120\binom{n+3}{6},$$

or with showing that $\sigma_3^m = \sigma_1^{2m}$ is the only identity of the form

$$\sigma_{i_1} \dots \sigma_{i_r} = \sigma_{j_1} \dots \sigma_{j_s}.$$

For some of the history of the subject, and for a selection of these articles, we mention **[1]**, **[3]**, **[5]**, **[7]**, **[9]**, **[11]**, **[12]**, **[13]** and **[16]**, and especially **[6]**, **[8]** and **[10]**.

Here, we shall take a quite different approach and generalise (1.1) to the extent that we describe *all polynomial relations that exist between any two of the* σ_i. As (1.1) simply asserts that the points $(\sigma_1(n), \sigma_3(n))$, $n = 1, 2, \dots$, lie on the parabola $y = x^2$, we are led naturally to (elementary) methods of algebraic geometry. The set of points (x, y) satisfying $T(x, y) = 0$, where T is a polynomial in two real variables, is a plane algebraic curve so that, writing

$$\Sigma_{ij} = \{(\sigma_i(n), \sigma_j(n)) \colon n = 1, 2, \dots\}, \tag{1.3}$$

the problem is to find all plane algebraic curves that contain the set Σ_{ij}.

It is well known that

$$\sigma_1 = \frac{n(n+1)}{2}, \qquad \sigma_2 = \frac{n(n+1)(2n+1)}{6}, \quad \sigma_3 = \frac{n^2(n+1)^2}{4},$$

and (by eliminating n from these) it is easy to find a polynomial relation between

σ_1 and σ_2, and between σ_2 and σ_3; these relations are

$$T(\sigma_1, \sigma_2) = 0, \qquad T(x, y) = 8x^3 + x^2 - 9y^2, \tag{1.4}$$

and

$$T(\sigma_2, \sigma_3) = 0, \qquad T(x, y) = 81x^4 - 18x^2y + y^2 - 64y^3, \tag{1.5}$$

respectively. Other, less obvious, relations are

$$T(\sigma_3, \sigma_5) = 0, \qquad T(x, y) = 16x^3 - x^2 - 6xy - 9y^2 \tag{1.6}$$

and

$$T(\sigma_2, \sigma_4) = 0, \qquad T(x, y) = 972x^5 - 7x^3 - 90x^2y - 375xy^2 - 500y^3. \tag{1.7}$$

The technique of eliminating a parameter enables us to prove much more than this, and we show

(1) for each pair of integers i and j with $1 \le i < j$, there is a unique irreducible polynomial T_{ij} in two variables, with integer coefficients, such that $T_{ij}(\sigma_i, \sigma_j) = 0$. By considering rings and ideals of polynomials, we can also show

(2) T_{ij} is the primitive relation between σ_i and σ_j in the sense that all other relations between these are trivial consequences of this one, and

(3) there is a (finite) algorithm for constructing any particular T_{ij}.

We remark in passing that we shall also see that there is no such result for polynomial relations among three or more of the σ_k.

To illustrate the results just described, consider the sums σ_1 and σ_3, and suppose that T is a real polynomial in two variables such that, for each $n, T(\sigma_1(n), \sigma_3(n)) = 0$. As the polynomial $T(t, t^2)$ vanishes when $t = n(n+1)/2$, $n = 1, 2, \dots$, it is identically zero and so $T(x, y)$ vanishes on the parabola $y = x^2$. One can show that this forces T to have a factor $y - x^2$, whence any relation $T(\sigma_1, \sigma_3) = 0$ is a trivial consequence of the primitive relation $\sigma_3 = \sigma_1^2$.

This paper is written to be available to as wide a readership as possible. Some historical references are given (our earliest source dates back to 1615), but no attempt has been made to identify the original sources.

2. THE COEFFICIENTS OF σ_k. It is a fundamental fact that $\sigma_k(n)$ is a polynomial in n of degree $k+1$ and it is worthwhile to review Pascal's elementary proof of this (given in 1654). It is simply that

$$\begin{aligned}(n+1)^{k+1} - 1 &= \sum_{m=1}^{n} \left[(m+1)^{k+1} - m^{k+1}\right] \\ &= \sum_{m=1}^{n} \sum_{r=0}^{k} \binom{k+1}{r} m^r \\ &= \sum_{r=0}^{k} \binom{k+1}{r} \sigma_r(n),\end{aligned}$$

from which it follows (by induction) that $\sigma_k(n)$ is a polynomial in n of degree $k+1$. This means that we can now legitimately consider $\sigma_k(z)$ *as a polynomial in a complex variable* z and we shall soon see that $\sigma_k(0) = 0$.

The Bernoulli numbers B_n, $n \ge 0$, first appeared in the posthumous work *Ars Conjectandi* by Jakob Bernoulli in 1713, although they were known by Faulhaber much earlier than this (see **[6]** and Chapter 10 in **[8]**). They were introduced in order to provide an explicit formula for the coefficients of the polynomial σ_k, and

are defined inductively by the recurrence relation

$$B_0 = 1, \quad \sum_{j=0}^{m} \binom{m+1}{j} B_j = 0 \tag{2.1}$$

([**14**], p. 229); we then have

$$1^k + 2^k + \cdots + (n-1)^k = \frac{1}{k+1} \sum_{j=0}^{k} \binom{k+1}{j} B_j n^{k+1-j} \tag{2.2}$$

([**14**], p. 234).

A calculation using (2.1) shows that

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0$$

(in fact, $B_3 = B_5 = B_7 = \cdots = 0$), so that, if $k \geq 3$, then

$$1^k + 2^k + \cdots + (n-1)^k = \frac{n^{k+1}}{k+1} - \frac{n^k}{2} + \frac{kn^{k-1}}{12} + O(n^{k-3}),$$

where here, $O(n^t)$ denotes a polynomial of degree at most t. Adding n^k to both sides, we obtain

$$\sigma_k(n) = \frac{n^{k+1}}{k+1} + \frac{n^k}{2} + \frac{kn^{k-1}}{12} + O(n^{k-3}); \tag{2.3}$$

note that *there is no term in* n^{k-2} (a consequence of $B_3 = 0$); this will be used later.

Of course, (2.2) shows that, for any complex number z,

$$\sigma_k(z) = \frac{1}{k+1} \sum_{j=0}^{k} \binom{k+1}{j} B_j (z+1)^{k+1-j}$$

which, with (2.1), yields

$$\sigma_k(0) = 0. \tag{2.4}$$

As an application of this, suppose that $T(\sigma_i, \sigma_j) = 0$ is any polynomial relation between σ_i and σ_j. Then, from (2.4),

$$0 = T(\sigma_i(0), \sigma_j(0)) = T(0,0);$$

thus *the constant term in T is zero*. Also, by putting $n = 1$, we see that *the sum of the coefficients of T is zero* (this is a useful check on our arithmetic).

3. FAULHABER POLYNOMIALS. It is well known that $\sigma_k(n)$ is a polynomial in n of degree $k+1$, but it is less well known that, when k is odd, σ_k is a polynomial in σ_1 of degree $\frac{1}{2}(k+1)$. The simplest case of this is $\sigma_3 = \sigma_1^2$, and the next two cases are

$$\sigma_5 = \sigma_1^2(4\sigma_1 - 1)/3, \quad \sigma_7 = \sigma_1^2(6\sigma_1^2 - 4\sigma_1 + 1)/3 \tag{3.1}$$

(proofs of these are given below). As these two formulae suggest, when k is odd, σ_1^2 divides σ_k. If k is even, then $\sigma_k(n)$ is of odd degree in n and so cannot be a polynomial in σ_1; however, in this case, σ_2 divides σ_k, and the quotient is a polynomial in σ_1. These results were known to Faulhaber and have been rediscovered many times since; even so, for the sake of completeness we give a formal statement and proof.

Theorem 3.1. *(i) For $k = 3, 5, \dots$ there exists a polynomial F_k, of degree $\frac{1}{2}(k+1)$ and with a double zero at the origin, such that $\sigma_k = F_k(\sigma_1)$.*

(ii) For $k = 2, 4, \dots$ there exists a polynomial F_k, of degree $\frac{1}{2}(k-2)$, such that $\sigma_k = \sigma_2 F_k(\sigma_1)$.

Proof: This is easy. Following the idea in Pascal's proof (namely, telescoping sums), we have

$$\begin{aligned}\sigma_1(n)^k &= \sum_{m=1}^{n}\left[\left(\frac{m(m+1)}{2}\right)^k - \left(\frac{(m-1)m}{2}\right)^k\right]\\ &= \sum_{m=1}^{n}\sum_{r=0}^{k}\binom{k}{r}\left(\frac{m}{2}\right)^k m^r\left[1-(-1)^{k-r}\right]\\ &= \frac{1}{2^k}\sum_{r=0}^{k}\binom{k}{r}\sigma_{k+r}(n)\left[1-(-1)^{k-r}\right].\end{aligned} \tag{3.2}$$

Now (3.2) holds for all k, but assume now that k is odd. Then the only terms in this sum that make a non-zero contribution are those with r even, and this show that σ_1^k is a linear combination of $\sigma_1, \sigma_3, \dots, \sigma_{2k-1}$. As $\sigma_3 = \sigma_1^2$, this provides the basis of a proof of (i) by induction; we omit the details.

We can prove (ii) in a similar way, so suppose now that k is even, and note that

$$\begin{aligned}(2n+1)\sigma_1(n)^k &= \sum_{m=1}^{n}\left[(2m+1)\left(\frac{m(m+1)}{2}\right)^k - (2m-1)\left(\frac{(m-1)m}{2}\right)^k\right]\\ &= \sum_{m=1}^{n}\sum_{r=0}^{k}\binom{k}{r}\left(\frac{m}{2}\right)^k m^r\left[(2m+1)-(2m-1)(-1)^{k-r}\right]\\ &= \frac{1}{2^k}\sum_{r=0}^{k}\binom{k}{r}\Big(2\sigma_{k+r+1}(n)\left[1+(-1)^{k-r+1}\right]\\ &\qquad +\sigma_{k+r}\left[1+(-1)^{k-r}\right]\Big).\end{aligned} \tag{3.3}$$

In the sum on the right, only the terms σ_q, with q even, survive, so that $(2n+1)\sigma_1(n)^k$ is a linear combination of $\sigma_2, \sigma_4, \dots, \sigma_{2k}$. Again, this provides the basis of a proof of (ii) by induction.

Notice that taking k to be 3, and then 5, in (3.2), we obtain the two identities in (1.2). Next, eliminating σ_3 from (1.1) and (1.2), we obtain the first identity in (3.1). Finally, taking $k = 4$ in (3.2) gives

$$\sigma_1^4 = \frac{1}{2}\sigma_5 + \frac{1}{2}\sigma_7,$$

and eliminating σ_5 from this and the first identity in (3.1), we obtain the second identity in (3.1).

The first few of the polynomials F_k in Theorem 3.1 are

$$\begin{aligned}F_3(t) &= t^2,\\ F_4(t) &= (6t-1)/5,\\ F_5(t) &= t^2(4t-1)/3\\ F_6(t) &= (12t^2-6t+1)/7\\ F_7(t) &= t^2(6t^2-4t+1)/3.\end{aligned}$$

The formulae for F_3, F_5 and F_7 are restatements of (1.1) and (3.1); the formulae for F_4 and F_6 are obtained by putting $k = 2$ and $k = 3$ in (3.3). Theorem 3.1 was known to Faulhaber in 1615, and it is suggested in **[6]** that the polynomials F_j (strictly, a mild variant of these) are called the *Faulhaber polynomials*. For more details, see **[6]**, **[7]** and **[8]** but, briefly, these ideas date back to Johann Faulhaber (1631). Later contributions were made by Fermat (1636), Pascal (1654), Bernoulli (1713), Euler (1755) and Jacobi (1834).

Because of the polynomial relation (1.4) between σ_1 and σ_2^2, Theorem 3.1 has the following corollary.

Corollary 3.2. (*i*) *If k is odd, σ_k is a polynomial in σ_1;*
(*ii*) *if k is even, σ_k^2 is a polynomial in σ_1.*

The apparent lack of symmetry between the cases k odd and k even can be overcome by the substitution $y = x + 1/2$. Then

$$\sigma_1(x) = \frac{y^2 - 1/4}{2}, \qquad \sigma_2(x) = \frac{y(y^2 - 1/4)}{3},$$

and, more generally, we see from Theorem 3.1 that

(1) if k is odd then $\sigma_k(x)$ is an even function of $x + 1/2$;
(2) if k is even then $\sigma_k(x)$ is an odd function of $x + 1/2$.

Although we will have no use for the following formulae, we end this section by recording that the F_k can be defined by generating functions (see **[9]**):

$$\frac{\cosh\left[(x/2)\sqrt{1+8t}\right] - \cosh(x/2)}{2\sinh(x/2)} = \sum_{r=0}^{\infty} F_{2r+1}(t)\frac{x^{2r+1}}{(2r+1)!},$$

and

$$\frac{\sinh\left[(x/2)\sqrt{1+8t}\right]}{2\sqrt{1+8t}\sinh(x/2)} = \frac{1}{2} + \frac{1}{3}\sum_{r=1}^{\infty} tF_{2r}(t)\frac{x^{2r}}{(2r)!}.$$

4. THE EXISTENCE OF RELATIONS. Some algebraic curves are given parametrically by, say, $x = f(t)$ and $y = g(t)$, where f and g are polynomials, and the technique of eliminating t to obtain the polynomial relation between x and y is sometimes referred to as the theory of elimination (see, for example, **[2]** pp. 179–181 and **[4]**, Chapter 3). As each $\sigma_k(n)$ is a polynomial in n, we can apply this theory to obtain a polynomial relation between any two of the σ_i. Indeed, this was the way we produced the relations (1.4) and (1.5) between σ_1 and σ_2, and between σ_2 and σ_3. The relation (1.6) between σ_3 and σ_5 can be obtained by eliminating not n but σ_1. According to Theorem 3.1 and the explicit expressions for the F_k, we have

$$\frac{\sigma_5}{\sigma_3} = \frac{4\sigma_1 - 1}{3},$$

whence

$$16\sigma_3 = (4\sigma_1)^2 = \left(\frac{3\sigma_5 + \sigma_3}{\sigma_3}\right)^2$$

which yields (1.6). Likewise, the relation (1.7) can be found directly by eliminating

σ_1 from the identities

$$\frac{\sigma_4}{\sigma_2} = \left(\frac{6\sigma_1 - 1}{5}\right), \qquad \sigma_2^2 = \frac{\sigma_1^2(8\sigma_1 + 1)}{9}.$$

To obtain a general result, we need a general theory and it is this that we now describe. Suppose that the two polynomials

$$f(x) = a_0 x^n + \cdots + a_{n-1}x + a_n, \qquad g(x) = b_0 x^m + \cdots + b_{m-1}x + b_m$$

have a common zero, say x_0. Then each of the equations

$$f(x) = xf(x) = \cdots = x^{m-1}f(x) = 0 = g(x) = xg(x) = \cdots = x^{n-1}g(x)$$

is of the form $p(x) = 0$, where p is a polynomial of degree (at most) $m + n - 1$, and, as each of these equations is satisfied when $x = x_0$, the determinant of the coefficients must vanish. This $(m + n) \times (m + n)$ determinant is the *resultant* $R(f,g)$ of f and g and, explicitly,

$$R(f,g) = \begin{vmatrix} a_0 & \cdots & \cdots & a_n & & \\ & \ddots & & & \ddots & \\ & & a_0 & \cdots & \cdots & a_n \\ b_0 & \cdots & \cdots & b_m & & \\ & \ddots & & & \ddots & \\ & & b_0 & \cdots & \cdots & b_m \end{vmatrix},$$

where the omitted elements are zero, and the diagonal of $R(f, g)$ contains m occurrences of a_0 and n of b_m. For more details, see, for example, **[2]**, **[4]** and **[15]**, pp. 83–88. For emphasis, we repeat that the existence of a common zero of f and g implies that $R(f, g) = 0$.

Let us now illustrate this use of the resultant by verifying the relation (1.4) between σ_1 and σ_2. As

$$2\sigma_1(n) = n^2 + n, \qquad 6\sigma_2(n) = n(n + 1)(2n + 1) = 2n^3 + 3n^2 + n,$$

the polynomials

$$f(t) = t^2 + t - 2\sigma_1(n), \qquad g(t) = 2t^3 + 3t^2 + t - 6\sigma_2(n)$$

have a common zero, namely, $t = n$. We deduce that, for each n,

$$\begin{vmatrix} 1 & 1 & -2\sigma_1(n) & 0 & 0 \\ 0 & 1 & 1 & -2\sigma_1(n) & 0 \\ 0 & 0 & 1 & 1 & -2\sigma_1(n) \\ 2 & 3 & 1 & -6\sigma_2(n) & 0 \\ 0 & 2 & 3 & 1 & -6\sigma_2(n) \end{vmatrix} = 0$$

and this simplifies to give (1.4).

A similar argument holds for any pair σ_i and σ_j so there is at least one non-trivial polynomial P with $P(\sigma_i, \sigma_j) = 0$. Pascal's argument shows that the coefficients of n in the polynomial σ_k are rational numbers; hence, after clearing denominators, we may assume that the coefficients of P are integers. Even more is true, for this argument shows that P is of degree $j + 1$ in σ_i and $i + 1$ in σ_j, and (by considering the expansion of the determinant) we even see that the term involving σ_i^{j+1} is independent of σ_j, and likewise with i and j interchanged; for

example, the determinant above is of the form

$$\begin{vmatrix} 2 & 3 \\ 0 & 2 \end{vmatrix} (-2\sigma_1)^3 + \cdots + \begin{vmatrix} 1 & 1 & -2\sigma_1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix} (-6\sigma_2)^2.$$

We have now proved the following result.

Proposition 4.1. *Given integers i and j with $1 \le i < j$, there is a polynomial $P(x, y)$, with integer coefficients and zero constant term, and of degree $j + 1$ in x and $i + 1$ in y with the leading terms in x and y having constant coefficients, such that $P(\sigma_i, \sigma_j) = 0$.*

This is not the end of the story, however. If we use this method in the case $i = 3$ and $j = 5$, we obtain a polynomial $P(x, y)$, of degree 6 in σ_3 and 4 in σ_5, such that $P(\sigma_3, \sigma_5) = 0$. It is tiresome (but instructive) to find this polynomial P explicitly, but in any event it is more complicated than the known relation (1.6) between σ_3 and σ_5 of *lower* degree. Clearly, the resultant obtained by eliminating n from the expressions $\sigma_3(n)$ and $\sigma_5(n)$ gives rise to a 10×10 determinant. If, on the other hand, we write σ_3 and σ_5 as polynomials in σ_1, and then eliminate σ_1, we obtain a 5×5 determinant which, after simplification, yields (1.6). The same reasoning applies to any pair σ_i and σ_j, where i and j are both odd integers, so that, in this case, it is better to eliminate σ_1 rather than n.

Now suppose that i and j are both even. We can write σ_i^2 and σ_j^2 as polynomials in σ_1 and then use the resultant to obtain a relation between σ_i^2 and σ_j^2 expressed as an $(i + j + 2) \times (i + j + 2)$ determinant which will be of degree $2(j + 1)$ in σ_i and $2(i + 1)$ in σ_j. If, on the other hand, we eliminate n between the expressions for σ_i and σ_j in terms of n, we obtain a determinant of the same size but with entries involving σ_i and σ_j (instead of σ_i^2 and σ_j^2). To illustrate this, observe that, using (2.2),

$$\sigma_2 = \frac{2n^3 + 3n^2 + n}{6}, \sigma_4 = \frac{6n^5 + 15n^4 + 10n^3 - n}{30},$$

whereas, using F_4 and (1.4),

$$\sigma_2^2 = \frac{8\sigma_1^3 + \sigma_1^2}{9}, \sigma_4^2 = \left(\frac{8\sigma_1^3 + \sigma_1^2}{9}\right)\left(\frac{6\sigma_1 - 1}{5}\right)^2.$$

In the case when i and j are both even, then, it is clearly better to eliminate the variable n rather than σ_1. We leave the reader to consider the case when i and j have opposite parity.

5. LÜROTH'S THEOREM AND THE RESULTANT.

Consider again the case when i and j are odd integers. We can either use the resultant to eliminate n and so obtain a relation $R(\sigma_i, \sigma_j) = 0$ between σ_i and σ_j, or we can express both as a function of σ_1 and then use the resultant to eliminate σ_1 and so obtain a second relation $R^*(\sigma_i, \sigma_j) = 0$ between σ_i and σ_j. In this section we shall describe the precise relationship between the two relations $R(\sigma_i, \sigma_j)$ and $R^*(\sigma_i, \sigma_j)$. The material in this section is related to Lüroth's Theorem ([**15**], pp. 198–200). The interested reader can consult [**15**] for a precise statement of this, but it is not necessary for it suffices to give the geometric interpretation described in [**15**]. Suppose that an algebraic curve is parametrised by $x = f(t)$ and $y = g(t)$, where f and g are polynomials. If each point of the curve corresponds to, say, d values of

t, then Lüroth's Theorem guarantees that there is a polynomial ϕ of degree d, and polynomials f_1 and g_1, such that

$$x = f(t) = f_1(\phi(t)), y = g(t) = g_1(\phi(t)); \tag{5.1}$$

it follows that we can take $s = \phi(t)$ as a new variable and parametrise the curve by the polynomials $x = f_1(s), y = g_1(s)$ of lower degree. In these circumstances, we can find the equation of the algebraic curve either by eliminating t, or by eliminating s. The two equations (arising from the resultants) are denoted by $R(f, g)$ and $R(f_1, g_1)$, respectively, and the relation between these is given in the following result.

Theorem 5.1. *In the above notation,* $R(f, g) = cR(f_1, g_1)^d$ *for some constant* c.

In our earlier discussion, we have expressed σ_i both as a polynomial in n and as a polynomial in σ_1, and the above discussion applies with $d = 2$ and

$$\phi(t) = \sigma_1(t) = t(t+1)/2.$$

We deduce that if i and j are odd integers, the resultant obtained by eliminating n is simply a scalar multiple of the square of the resultant obtained by eliminating σ_1.

The proof of Theorem 5.1. We shall work with complex numbers (so that all polynomials factorise into linear factors). Suppose first that f and g are any complex polynomials, say

$$f(z) = a(z - z_1) \cdots (z - z_n), \qquad g(z) = b(z - w_1) \cdots (z - w_m).$$

As each coefficient of f is the product of a with a symmetric function of the roots z_j, and similarly for g, the resultant $R(f, g)$ is of the form $a^m b^n P(z_1, \dots, z_n, w_1, \dots, w_m)$ for some polynomial P. Thinking of the roots z_i and w_j as variables, we note that if $z_i - w_j = 0$, then f and g have a common root and so $R(f, g) = 0$. Continuing this line of argument (the details can be found in **[15]**, p. 86), we find that

$$R(f, g) = a^m b^n \prod_{i,j} (z_i - w_j).$$

We shall now apply these ideas to prove Theorem 5.1.

Let f, g, f_1, g_1 and ϕ now be the polynomials in (5.1) and suppose that these have degrees nd, md, n, m and d, respectively. We denote the zeros of f_1 by $\alpha_1, \dots, \alpha_n$, and the solutions of $\phi(z) = \alpha_j$ by $x_{1j}, \dots, x_{dj}$; then the zeros of f are precisely the numbers x_{ij}, where $i = 1, \dots, d$ and $j = 1, \dots, n$. Likewise, we denote the zeros of g_1 by $\beta_1, \dots, \beta_m$, and the solutions of $\phi(z) = \beta_s$ by $y_{1s}, \dots, y_{ds}$; then the zeros of g are the numbers y_{rs}. Finally, we use A_1, A_2, A_3, A_4 to denote constants (which we do not bother to evaluate).

According to the first paragraph in the proof,

$$R(f, g) = A_1 \prod_{i,j,r,s} (x_{ij} - y_{rs}), \quad R(f_1, g_1) = A_2 \prod_{j,s} (\alpha_j - \beta_s).$$

However, we also have

$$\phi(z) - \beta_s = A_3(z - y_{1s}) \cdots (z - y_{ds}) = A_3 \prod_r (z - y_{rs}),$$

so that, for each i, $1 \le i \le d$,

$$\alpha_j - \beta_s = \phi(x_{ij}) - \beta_s = A_3 \prod_r (x_{ij} - y_{rs}).$$

This holds for each i so, taking the product of both sides over $i = 1, \dots, d$, we obtain

$$(\alpha_j - \beta_s)^d = A_3^d \prod_{i,r} (x_{ij} - y_{rs})$$

from which the result follows immediately.

6. IDEALS OF POLYNOMIALS. In this section we characterise, for a given pair i and j satisfying $1 \le i < j$, the totality of polynomials T with integer coefficients for which $T(\sigma_i, \sigma_j) = 0$. To achieve this we borrow an idea from algebraic geometry and study the family of polynomials that vanish on a given set; for more details, we recommend [**4**]. We denote the integers by $\mathbf{Z}$, and the class (or ring) of polynomials with integer coefficients and in the two variables x and y by $\mathbf{Z}[x, y]$. The key fact that we need about $\mathbf{Z}[x, y]$ is that it is a unique factorisation domain; this means that any polynomial in $\mathbf{Z}[x, y]$ may be factored into a product of irreducible polynomials, and that up to order and factors of -1, this factorisation is unique (see [**2**], pp. 172–176).

Our declared aim is to study the family

$$\Gamma = \{T \in \mathbf{Z}[x, y] \colon T(\sigma_i, \sigma_j) = 0\}, \tag{6.1}$$

which we prefer to write as

$$\Gamma = \{T \in \mathbf{Z}[x, y] \colon T = 0 \text{ on } \Sigma_{ij}\},$$

where Σ_{ij} is given in (1.3) as the set of points $\{(\sigma_i(n), \sigma_j(n)) \in \mathbb{R}^2 \colon n = 1, 2, \dots\}$.

For the moment, let E be any non-empty subset of $\mathbb{R}^2$, and define

$$\mathscr{I}(E) = \{T \in \mathbf{Z}[x, y] \colon T = 0 \text{ on } E\}.$$

The set $\mathscr{I}(E)$ is known in ring theory as an *ideal*, for it is closed under addition and the product $T_1 T_2$ is in $\mathscr{I}(E)$ whenever one of the T_i is. We want to investigate circumstances under which $\mathscr{I}(E)$ consists of all polynomial multiples of a single polynomial $T_0(x, y)$ (then $\mathscr{I}(E)$ is said to be the *principal ideal* generated by T_0). In general, this will not be so; it is not when, for example, E is the intersection of the two co-ordinate axes. However, we do have the following result.

Lemma 6.1. *Let f and g be non-constant polynomials with rational coefficients in one real variable, and let $E = \{(f(n), g(n)) \colon n = 1, 2, \dots\}$. Then $\mathscr{I}(E)$ is generated by a non-trivial irreducible polynomial in $\mathbf{Z}[x, y]$.*

Proof: We have already seen that there is a polynomial P in two variables such that for all integers n, $P(f(n), g(n)) = 0$. This P can be obtained from the resultant of f and g (as in the proof of Proposition 4.1) so that if f and g have rational coefficients P may be taken to have integer coefficients. We claim that we may also assume that P is irreducible, for suppose that $P = P_1 \dots P_l$, where the P_j are the irreducible factors of P. Then there must be (at least) one factor P_j such that the polynomial $P_j(f(n), g(n))$ vanishes for infinitely many integer values of n. It follows that the polynomial $P_j(f(x), g(x))$ has infinitely many zeros and so vanishes for all x, and hence for all integers n. In conclusion, there is a irreducible polynomial P in $\mathbf{Z}[x, y]$ such that $P(f, g) = 0$.

We shall now show that P divides any polynomial T in $\mathbf{Z}[x, y]$ for which $T(f, g) = 0$. First, we express P and T as polynomials in y (whose coefficients are polynomials in x) and we then compute the resultant of P and T by eliminating the variable y. The resultant is a polynomial $R(x)$, and it is known that there are polynomials A and B in $\mathbf{Z}[x, y]$ such that

$$A(x,y)P(x,y) + B(x,y)T(x,y) = R(x),$$

where (as polynomials in y) $\deg(B) < \deg(P)$ and $\deg(A) < \deg(T)$ (see **[2]**, Proposition 4.2.4, p. 179, and p. 192). As $P(f(n), g(n))$ and $T(f(n), g(n))$ vanish for each integer n, we see that $R(n) = 0$ for each integer n. It follows that R is the zero polynomial, and hence $P(x, y)$ divides $B(x, y)T(x, y)$. As P is irreducible, it divides B or T, and as $\deg(B) < \deg(P)$, it cannot divide B. It follows that P divides T, and that $\mathscr{I}(E)$ is the principal ideal generated by P.

The discussion to date yields our main result which follows.

Theorem 6.2. *Let i and j be integers with $1 \le i < j$. Then there is a non-constant irreducible polynomial T_{ij} in $\mathbf{Z}[x, y]$ such that $T_{ij}(\sigma_i, \sigma_j) = 0$. Further, T_{ij} divides P for any P in $\mathbf{Z}[x, y]$ for which $P(\sigma_i, \sigma_j) = 0$.*

Of course, this result says that any polynomial relation between σ_i and σ_j is a trivial consequence of the relation $T_{ij}(\sigma_i, \sigma_j) = 0$; for example, $T_{13}(x, y) = y - x^2$ and so if $P(\sigma_1, \sigma_3) = 0$, then P contains a factor $y - x^2$ and so the relation holds because of the existence of this factor. If P is any non-constant irreducible polynomial in $\mathbf{Z}[x, y]$, and if $P(\sigma_i, \sigma_j) = 0$, then $P = \pm T_{ij}$ and this observation enable us to identify T_{ij} in certain cases. For example, to show that

$$T_{23}(x,y) = 81x^4 - 18x^2y + y^2 - 64y^3, \qquad T_{35}(x,y) = 16x^3 - x^2 - 6xy - 9y^2,$$

we have, because of (1.5) and (1.6), only to prove that these polynomials are irreducible over $\mathbf{Z}$ and this can easily be done by assuming the contrary and reaching a contradiction. Finally, Faulhaber's observations mean that, for odd k, $T_{1k}(x, y)$ is an integer multiple of $y - F_k(x)$, whereas for even k, $T_{1k}(x, y)$ is an integer multiple of $y^2 - x^2(8x + 1)F_k(x)^2$.

Using the resultant we can eliminate n from $\sigma_i(n)$ and $\sigma_j(n)$ (in a finite number of steps) and obtain an explicit polynomial P in $\mathbf{Z}[x, y]$ for which $P(\sigma_i, \sigma_j) = 0$. This P must contain T_{ij} as a factor, and since the factorisation of P (over $\mathbf{Z}$) can also be carried out in a finite number of steps (see **[15]**, p. 77), it follows that *each T_{ij} is computable in a finite number of steps*. There are, of course, other ways of finding the T_{ij}, for example, by using the Groebner basis method described in **[4]**. In the (few) examples I have tried with i and j odd, the polynomial relation between σ_i and σ_j obtained by eliminating σ_1 is irreducible over $\mathbf{Z}$, and hence is T_{ij}. I have been unable to prove that this is true for all odd i and j, but if it were to be true, it would provide a beautiful relationship between the ideal of polynomials annihilating (σ_i, σ_j) and Faulhaber's contribution 350 years ago.

7. SEPARABILITY. There is one feature of the relations between the σ_i that we have not yet commented on. We say that the relation $T_{ij}(\sigma_i, \sigma_j) = 0$, where $1 \le i < j$, is *separable* if it can be expressed in the form

$$P(\sigma_i) = Q(\sigma_j), \tag{7.1}$$

where P and Q are polynomials in one variable. Faulhaber's results show that

every T_{ij} with $i = 1$ is separable, and all the examples we have of T_{ij} with $2 \le i < j$ are not separable. This suggests that T_{ij} is separable if and only if $i = 1$ and we shall now show that this is so. This means, of course, that the only cases in which T_{ij} is separable are those found by Faulhaber in the seventeenth century.

Theorem 7.1. *The relation $T_{ij}(\sigma_i, \sigma_j) = 0$ is separable if and only if $i = 1$.*

Proof: We assume that (7.1) holds with $2 \le i < j$ and we seek a contradiction. We write

$$P(x) = Ax^r + O(x^{r-1}), \qquad Q(x) = Bx^s + O(x^{s-1}), \tag{7.2}$$

where $AB \neq 0$, and, by comparing degrees of the two polynomials in (7.1), we have

$$(i+1)r = (j+1)s = N, \tag{7.3}$$

say. Observe that as $j \ge 3$, we have $N \ge 4$.

Now $Q(\sigma_j)$ is a polynomial of degree N in the variable n and, as

$$(j+1)(s-1) = N - (1+j) \le N - 4,$$

we have, from (2.3) and (7.2),

$$\begin{aligned} Q(\sigma_j(n)) &= B\left[\left(\frac{n^{j+1}}{j+1} + \frac{n^j}{2} + \frac{jn^{j-1}}{12}\right) + O(n^{j-3})\right]^s + O(n^{N-4}) \\ &= B\left(\frac{n^{j+1}}{j+1} + \frac{n^j}{2} + \frac{jn^{j-1}}{12}\right)^s + O(n^{N-4}) \end{aligned}$$

because $(j+1)(s-1) + (j-3) = N - 4$. The same holds for $P(\sigma_i)$; thus

$$A\left(\frac{n^{i+1}}{i+1} + \frac{n^i}{2} + \frac{in^{i-1}}{12}\right)^r = B\left(\frac{n^{j+1}}{j+1} + \frac{n^j}{2} + \frac{jn^{j-1}}{12}\right)^s + O(n^{N-4}).$$

Using (7.3), and equating the coefficients of x^N, we obtain

$$\frac{A}{(i+1)^r} = \frac{B}{(j+1)^s},$$

and this leads to

$$\begin{aligned} &\left(n^{i+1} + \frac{(i+1)n^i}{2} + \frac{i(i+1)n^{i-1}}{12}\right)^r \\ &\quad = \left(n^{j+1} + \frac{(j+1)n^j}{2} + \frac{j(j+1)n^{j-1}}{12}\right)^s + O(n^{N-4}). \end{aligned}$$

Working now to an error term of order n^{N-3}, and using (7.3), this simplifies to

$$\begin{aligned} &24n^N + 12Nn^{N-1} + 2iNn^{N-2} + 3N(N-i-1)n^{N-2} \\ &\quad = 24n^N + 12Nn^{N-1} + 2jNn^{N-2} + 3N(N-j-1)n^{N-2}. \end{aligned}$$

As this implies that $i = j$ (contrary to our assumption) Theorem 7.1 is proved.

8. CONCLUDING REMARKS. The reader may have noticed that much of the above discussion does not depend on the particular nature of the polynomials

$\sigma_k(n)$ as sums of powers of integers. Indeed, if

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n, \qquad g(x) = b_0 + b_1 x + \cdots + b_m x^m,$$

are any real polynomials, then the polynomials

$$F(t) = [a_0 - f(y)] + a_1 t + \cdots + a_n t^n, \quad G(t) = [b_0 - g(y)] + b_1 t + \cdots + b_m t^m,$$

have a common zero, namely $t = y$, and so $R(F, G) = 0$. This means, of course, that there is some polynomial $T(x, y)$ such that $T(f, g) = 0$.

It is amusing to apply these ideas to the Tchebychev polynomials T_n, which are defined by

$$T_n(\cos\theta) = \cos(n\theta).$$

As

$$T_n(T_m(\cos\theta)) = \cos(mn\theta) = T_m(T_n(\cos\theta)),$$

we see that $T_n(T_m(t)) = T_m(T_n(t))$ for all t; thus the polynomial relation connecting T_m and T_n is

$$T(T_m, T_n) = 0, \qquad T(x, y) = T_n(x) - T_m(y).$$

In this case, then, all of the primitive relations are separable.

As a final example, consider the Legendre polynomials ϕ_n defined by $\phi_0(x) = 1$ and

$$\phi_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}\left([1 - x^2]^n\right).$$

Now

$$\phi_2(x) = \frac{3x^2 - 1}{2}, \qquad \phi_3(x) = \frac{5x^3 - 3x}{2},$$

so that $4\phi_3(x)^2$ is polynomial in x^2 and $x^2 = (2\phi_2(x) + 1)/3$. This leads directly to the relation

$$T(\phi_2, \phi_3) = 0, \quad T(x, y) = 50x^3 - 15x^2 - 12x + 4 - 27y^2,$$

a result that can also be obtained by using the resultant.

REFERENCES

1. Anderson, I., Sums of squares and binomial coefficients, *Math. Gazette*, 65 (1981), 87–92.
2. Brieskorn, E. and Knörrer, H., *Plane algebraic curves*, Birkhäuser, 1986.
3. Colledge, T., *Pascal's triangle*, Stradbroke: Tarquin, 1992.
4. Cox, D., Little, J. and O'Shea, D., *Ideals, varieties and algorithms*, Springer-Verlag, 1992.
5. Edmonds, S. M., Sums of powers of the natural numbers, *Mathematical Gazette*, 41 (1957), 187–189.
6. Edwards, A. W. F., Sums of powers of integers: a little of the history, *Math. Gazette*, 66 (1982), 22–29.
7. Edwards, A. W. F., A quick route to sums of powers, *Amer. Math. Monthly*, 93 (1986), 451–455.
8. Edwards, A. W. F., *Pascal's arithmetical triangle*, Charles Griffin & Co., 1987.
9. Gessel, I., A formula for power sums, *Amer. Math. Monthly*, 95 (1988), 961–962.
10. Knuth, D. E., Johann Faulhaber and sums of powers, *Mathematics of Computation*, 61 (1993), 277–294.
11. Parker, R. V., Sums of powers of the integers, *Mathematical Gazette*, 42 (1958), 91–95.
12. Penney, D. E. and Pomerance, C., Multiplicative relations for sums of initial k-th powers, *Amer. Math. Monthly*, 92 (1985), 729–731.

13. Piza, P. A., Powers of sums and sums of powers, *Mathematics Magazine*, 25 (1952), 137–142.
14. Rose, H. E., *A course in number theory*, Oxford Univ. Press, 1989.
15. Van de Waerden, B. L., *Modern Algebra*, Volume 1, Ungar Pub. Co., 1949.
16. Woodall, D. R., Finite sums, matrices and induction, *Math. Gazette*, 65 (1981), 92–103.

Department of Pure Mathematics & Mathematical Statistics
University of Cambridge
16 Mill Lane
Cambridge CB2 1SB
ENGLAND
a.f.beardon@pmms.cam.ac.uk

PICTURE PUZZLE

(from the collection of Paul Halmos)

The picture was taken in 1980. Could you recognize him without the beard?
(see page 229)

6. CONCLUDING REMARKS. As I remarked in the introduction, there's still much more we could do. Discussions about Lie groups, group actions and fiber bundles follow naturally from the material presented in this paper, as does a more in-depth treatment of the geometry of S^3. But for this exposition, I think that now is the right time to say 'Enough!'. I hope that this article has provided you with a better understanding (and a usable model) of the 3-sphere, and that you are inspired to pursue further some of the ideas introduced above.

REFERENCES

1. S. Akbulut and J. McCarthy, *Casson's Invariant for Oriented Homology 3-Spheres An Exposition*, Mathematical Notes 36, Princeton University Press, Princeton, 1990.
2. M. Artin, *Algebra*, Prentice Hall, Englewood Cliffs, New Jersey, 1991.
3. C. Chevalley, *Theory of Lie Groups, I*, Princeton University Press, Princeton, 1946.
4. A. Knapp, *Lie Groups, Lie Algebras, and Cohomology*, Mathematical Notes 34, Princeton University Press, Princeton, 1988.
5. S. MacLane and G. Birkhoff, *Algebra*, 2nd ed., Collier Macmillan, New York, 1979.

Department of Mathematics
Rice University
P.O. Box 1892
Houston, TX 77251
zulli@math.rice.edu

A Ph.D. dissertation is a paper of the professor written under aggravating circumstances.

Attributed by G. Pòlya to Adolph Hurwitz, as quoted in Howard W. Eves, *Mathematical Circles Revisited*, Prindle, Weber and Schmidt, Boston, 1971.

Answer to Picture Puzzle

(p. 213)

Raoul Bott

Wielandt's Theorem About the Γ-function

Reinhold Remmert

It is well known that the Γ-function cannot be characterised by its functional equation $\Gamma(z+1) = z\Gamma(z)$ and the condition $\Gamma(1) = 1$. In 1922 H. BOHR and J. MOLLERUP showed in **[BM]** that the additional assumption of *logarithmic convexity* yields the uniqueness of $\Gamma(x)$ for real $x > 0$. Everyone admires Emil ARTIN's treatise **[A]** from 1931 with its beautiful applications of the BOHR-MOLLERUP theorem. It is hardly known that there is also an elegant *function theoretic characterization of* $\Gamma(z)$. This uniqueness theorem was discovered by Helmut WIELANDT in 1939 and is at the centre of this note. A function theorist ought to be as much fascinated by WIELANDT's complex-analytic characterization as by the BOHR-MOLLERUP theorem. WIELANDT's theorem immediately yields classical results about the Γ-function; as examples we shall derive

- the GAUSS product from the EULER integral,
- the multiplication formulae of GAUSS,
- the representation of the Beta function by Gamma functions,
- STIRLING's formula.

1. THE FUNCTIONAL EQUATION. We consider holomorphic functions f in the *right half plane* $\mathbb{A} := \{z \in \mathbb{C}\colon \operatorname{Re} z > 0\}$ satisfying the equation

$$f(z+1) = zf(z) \textit{ for all points } z \in \mathbb{A}. \tag{1}$$

By induction we obtain for all $n \in \mathbb{N} := \{0, 1, 2, 3, \dots\}$ and all $z \in \mathbb{A}$

$$f(z+n+1) = z(z+1)\cdots(z+n)f(z). \tag{2}$$

Now it is easily shown:

Every function f holomorphic in $\mathbb{A}$ and satisfying (1) *admits a meromorphic extension $\hat{f}$ to $\mathbb{C}$. This function $\hat{f}$ is holomorphic in $\mathbb{C} \setminus \{0, -1, -2, -3, \dots\}$; the point $-n$ is a pole of order ≤ 1 with residue $(-1)^n f(1)/n!$, where $n \in \mathbb{N}$.* (3)

In particular $\hat{f}$ is an entire function if and only if $f(1) = 0$.

Proof: Take a point $z \in \mathbb{C}$ such that $-z \notin \mathbb{N}$. Then $\hat{z} := z + n + 1 \in \mathbb{A}$ for large $n \in \mathbb{N}$ and we may define $\hat{f}(z) := f(\hat{z})/[z(z+1)\cdots(z+n)] \in \mathbb{C}$. Clearly this definition is independent of the choice of n and we get a holomorphic function $\hat{f}$ in $\mathbb{C} \setminus \{0, -1, -2, -3, \dots\}$ such that $\hat{f}|\mathbb{A} = f$. Furthermore

$$\lim_{z \to -n} (z+n)\hat{f}(z) = \lim_{z \to -n} \frac{f(z+n+1)}{z(z+1)\cdots(z+n-1)} = \frac{(-1)^n}{n!} f(1)$$

for all $n \in \mathbb{N}$. This shows that $-n$ is a pole of $\hat{f}$ of order ≤ 1 with the residue $(-1)^n f(1)/n!$.

2. EULER'S INTEGRAL. Our point of departure is the equation

$$\Gamma(z) := \int_0^{\infty} t^{z-1} e^{-t}\, dt,\ z \in \mathbb{A}. \tag{1}$$

The integral converges *uniformly* in every strip $\{z \in \mathbb{C}\colon a \leq \operatorname{Re} z \leq b\}$ with $0 < a < b < \infty$; for a proof we refer to classical books. By WEIERSTRASS's convergence theorem the limit function $\Gamma(z)$ is holomorphic in $\mathbb{A}$. Partial integration is legitimate and yields at once

$$\Gamma(z+1) = z\Gamma(z) \quad \textit{for all } z \in \mathbb{A},\ \Gamma(1) = 1. \tag{2}$$

Furthermore the equation $|t^{z-1}| = t^{\operatorname{Re} z - 1}$ entails directly

$$|\Gamma(z)| \leq |\Gamma(\operatorname{Re} z)| \quad \textit{for all } z \in \mathbb{A}; \tag{3}$$

in particular $\Gamma(z)$ is *bounded in every strip* $\{z \in \mathbb{C}\colon a \leq \operatorname{Re} z \leq b\}$, $0 < a < b < \infty$.

3. WIELANDT'S THEOREM. *Let $F(z)$ be a holomorphic function in the right half plane $\mathbb{A}$ having the following two properties*:

(a) $F(z+1) = zF(z)$ *for all* $z \in \mathbb{A}$.

(b) $F(z)$ *is bounded in the strip* $S := \{z \in \mathbb{C}\colon 1 \leq \operatorname{Re} z < 2\}$.

Then $F(z) = a\Gamma(z)$ in $\mathbb{A}$ with $a := F(1)$.

Proof: The function $f := F - a\Gamma$ is holomorphic in $\mathbb{A}$. From a) we obtain: $f(z+1) = zf(z)$ for all $z \in \mathbb{A}$. Since $f(1) = 0$, we conclude from 1(3) that f extends to an entire function $\hat{f}$. Since $\Gamma|S$ is bounded by 2(3), the function $f|S$ is bounded by (b). This implies *boundedness of f on* $S_0 := \{z \in \mathbb{C}\colon 0 \leq \operatorname{Re} z < 1\}$: For all points $z \in S_0$ with $|\operatorname{Im} z| \leq 1$ this is clear; for all points with $|\operatorname{Im} z| > 1$ this follows from $f(z) = f(z+1)/z$ and the boundedness of f on S.

We now consider the entire function $s(z) := \hat{f}(z)\hat{f}(1-z)$. Since $f(z)$ and $\hat{f}(1-z)$ take the same values in S_0 the function $s(z)$ is bounded in S_0. Since $\hat{f}(z+1) = z\hat{f}(z)$ and $\hat{f}(-z) = -\hat{f}(1-z)/z$ we see: $s(z+1) = -s(z)$. Hence $s(z)$ is bounded in $\mathbb{C}$ and therefore constant by LIOUVILLE's theorem. We get $s(z) \equiv s(1) = f(1)\hat{f}(0) = 0$. This implies $\hat{f}(z) \equiv 0$, i.e. $F(z) = a\Gamma(z)$ in $\mathbb{A}$.

4. HISTORICAL NOTE. In 1941, in the fifth edition of his classical booklet *Funktionentheorie II*, Konrad KNOPP presents, on pages 47 to 49, WIELANDT's uniqueness theorem; in a footnote on page 49 he writes: "Diesen Beweis verdanke ich H. WIELANDT." [This proof I owe to H. WIELANDT]. In a letter dated March 5, 1990, WIELANDT wrote to me: "Als ich 1938 die Bibliothek des Mathematischen Instituts in Tübingen zu verwalten hatte, habe ich der Funktionentheorie gefrönt neben meinem Spezialgebiet, der Gruppentheorie. Wenn mir etwas einfiel, was vielleicht für künftige Vorlesungen brauchbar sein könnte, habe ich es gelegentlich in vortragsfertiger Form notiert, und Knopp gezeigt, soweit es für eine Neuauflage seiner für mich so anregenden Göschenbändchen von Interesse sein konnte." [When I was in charge of the Mathematics Institute library at Tübingen in 1938, I dabbled in function theory, along with my specialty, group theory. Whenever something occurred to me that could possibly be of value in future lecturing, I wrote it down in a form that was ready for use and showed it to Knopp, in case it might be of interest for any new edition of his little Goeschen paperback that I found so stimulating.] For further details see [**R**$_2$].

Already in 1914 G. D. BIRKHOFF had emphasized using the elements of the general theory of functions of a complex variable to obtain the principal properties of the Γ-function. He derived two fundamental identities by studying quotients of functions in the closed strip $\overline{S}$ and showing that the quotients are constant due to LIOUVILLE's theorem, cf. **[Bi]**, pp. 8 and 10. Was he close to a uniqueness theorem for $\Gamma(z)$?

5. GAUSS PRODUCT AND WEIERSTRASS FORMULA. We put

$$\gamma := \lim_{n\to\infty}\left(\sum_{\nu=1}^{n}\frac{1}{\nu} - \log n\right) \simeq 0.577\ldots$$

and take for granted that

$$\Delta(z) := ze^{\gamma z}\prod_{\nu=1}^{\infty}\left(1+\frac{z}{\nu}\right)e^{-z/\nu} \tag{1}$$

converges uniformly on every compact subset of $\mathbb{C}$ to an entire function $\Delta(z)$. We write n^z for $e^{z\log n}$ and claim the GAUSS formula

$$\Gamma(z) = \frac{1}{\Delta(z)} = \lim_{n\to\infty}\frac{n!n^z}{z(z+1)\cdots(z+n)} \tag{2}$$
$$\textit{for all } z \in \mathbb{C}\setminus\{0,-1,-2,-3,\ldots\}.$$

Proof: Putting $G_n(z) := n!n^z/[z(z+1)\cdots(z+n)]$ we see that

$$G_n(z)\cdot ze^{\gamma z}\prod_{\nu=1}^{n}\left(1+\frac{z}{\nu}\right)e^{-z/\nu} = e^{\gamma z}e^{z\log n}\prod_{\nu=1}^{n}e^{-z/\nu} = e^{z(\gamma+\log n-\sum_{\nu=1}^{n}(1/\nu))}$$

for all $n \geq 1$ and all $z \in \mathbb{C}\setminus\{0,-1,-2,-3,\ldots\}$. We conclude

$$\lim_{n\to\infty}G_n(z) = 1/\Delta(z) \text{ for all } z \in \mathbb{C}\setminus\{0,-1,-2,-3,\ldots\}.$$

The function $G(z) := 1/\Delta(z)$ is holomorphic in $\mathbb{C}\setminus\{0,-1,-2,-3,\ldots\}$. Since $G_n(1) = n/(n+1)$ and $G_n(z+1) = zG_n(z)[n/(z+n+1)]$, we have $G(1) = 1$ and $G(z+1) = zG(z)$. Since $|n^z| = n^x$ and $|z+\nu| \geq x+\nu$ for all z with $x :=$ Re $z > 0$, we get $|G_n(z)| \leq G_n(x)$ for all $z \in \mathbb{A}$. Hence $|G(z)| \leq G(x)$ is bounded in the strip S. Therefore $G(z) = \Gamma(z)$ by WIELANDT. q.e.d.

Combining (1) and (2) we obtain the WEIERSTRASS formula

$$1/\Gamma(z) = ze^{\gamma z}\prod_{\nu=1}^{\infty}\left(1+\frac{z}{\nu}\right)e^{-z/\nu} \text{ for all } z \in \mathbb{C}\setminus\{0,-1,-2,-3,\ldots\}. \tag{3}$$

Since $\Gamma(1-z) = -z\Gamma(-z)$ and $\sin\pi z = \pi z\prod_{\nu=1}^{\infty}(1-z^2/\nu^2)$, we get from (3)

$$\Gamma(z)\Gamma(1-z) = \pi/\sin\pi z. \tag{4}$$

Using the equation $2^{n-1}\prod_1^{n-1}\sin(\nu\pi/n) = n$, we derive EULER's formula

$$\sqrt{n}\,\Gamma\left(\frac{1}{n}\right)\Gamma\left(\frac{2}{n}\right)\cdots\Gamma\left(\frac{n-1}{n}\right) = (2\pi)^{(n-1)/2}, \qquad n = 2,3,\ldots. \tag{5}$$

[The product $\prod_1^{n-1}\sin(\nu\pi/n)$ is computed as follows: Since $2i\sin z = e^{iz}(1-e^{-2iz})$ and $\prod_1^{n-1}e^{i\nu\pi/n} = i^{n-1}$ we see $(2i)^{n-1}\prod_1^{n-1}\sin(\nu\pi/n) = i^{n-1}\prod_1^{n-1}(1-e^{-2i\nu\pi/n})$. The last product is the value of $(w^n-1)/(w-1)$ at 1, i.e. the number n].

6. MULTIPLICATION FORMULAE OF GAUSS. We want to show

$$\prod_{\nu=0}^{n-1}\Gamma\left(z+\frac{\nu}{n}\right)=(2\pi)^{(n-1)/2}n^{1/2-nz}\Gamma(nz),\qquad n=2,3,\dots;\ z\in\mathbb{A}.\quad(1)$$

Proof: Fix such an n. The function

$$F(z):=\Gamma\left(\frac{z}{n}\right)\Gamma\left(\frac{z+1}{n}\right)\cdots\Gamma\left(\frac{z+n-1}{n}\right)\Big/(2\pi)^{(n-1)/2}n^{1/2-z}$$

is holomorphic in $\mathbb{A}$. Using 2(2) we see:

$$F(z+1)=\Gamma\left(\frac{z}{n}\right)^{-1}F(z)\Gamma\left(\frac{z+n}{n}\right)\Big/n^{-1}=zF(z)\quad\text{for all } z\in\mathbb{A}.$$

Furthermore EULER's formulae 5(5) says $F(1)=1$. Since $\Gamma(z)$ is bounded in the strip S the function $F(z)$ is also bounded in S. Hence $F=\Gamma$ by WIELANDT's theorem. q.e.d.

GAUSS proved (1) in 1812, cf. [G], p. 150. For $n=2$ we have the *duplication formula of* LEGENDRE:

$$\sqrt{\pi}\,\Gamma(2z)=2^{2z-1}\Gamma(z)\Gamma\left(z+\frac{1}{2}\right).\quad(2)$$

7. BETA FUNCTION. This function of two complex variables w, z is defined by

$$B(w,z):=\int_0^1 t^{w-1}(1-t)^{z-1}\,dt,\qquad w,z\in\mathbb{A}.\quad(1)$$

The integral converges uniformly on compact sets in $\mathbb{A}\times\mathbb{A}$ and hence defines a *holomorphic* function in $\mathbb{A}\times\mathbb{A}$. From (1) we get at once:

(a) $$B(w,1)=w^{-1},\quad B(w,z+1)=\frac{z}{w+z}B(w,z),$$

(b) $$|B(w,z)|\le B(\operatorname{Re}w,\operatorname{Re}z).$$

The second formula in (a) holds since

$$\begin{aligned}&(w+z)B(w,z+1)-zB(w,z)\\&=(w+z)\int_0^1 t^{w-1}(1-t)^z\,dt-z\int_0^1 t^{w-1}(1-t)^{z-1}\,dt\\&=\int_0^1\left\{wt^{w-1}(1-t)^z-t^w z(1-t)^{z-1}\right\}dt\\&=\left[t^w(1-t)^z\right]_0^1=0.\end{aligned}$$

Already in 1771 EULER knew:

$$B(w,z)=\frac{\Gamma(w)\Gamma(z)}{\Gamma(w+z)}\quad\textit{for all } w,z\in\mathbb{A}.\quad(2)$$

Proof: Fix $w\in\mathbb{A}$. The function $F(z):=B(w,z)\Gamma(w+z)$ is holomorphic in $\mathbb{A}$. Clearly $F(1)=\Gamma(w)$ and $F(z+1)=zF(z)$ by a) and 2(2). From b) and the inequality $|\Gamma(w+z)|\le\Gamma(\operatorname{Re}(w+z))$ we conclude that $F(z)$ is bounded in the strip S. Hence we have $F(z)=\Gamma(w)\Gamma(z)$ by WIELANDT. q.e.d.

Let us mention in passing that R. DEDEKIND wrote his dissertation in 1852 on the functions $\Gamma(z)$ and $B(w, z)$, cf. [D]. This solid piece of work gave no indication that a star had been born.

8. THE ERROR FUNCTION $\mu(z)$. STIRLING'S FORMULA. The classical STIRLING formula says

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n}, \quad \text{that is, } \Gamma(n) \sim \sqrt{2\pi} n^{n-1/2} e^{-n}.$$

The last formula can be made more precise. We denote by $\mathbb{C}^-$ the *slit* plane $\mathbb{C} \setminus \{x \in \mathbb{R}: x \leq 0\}$ and claim:

Theorem. (STIELTJES 1889). *There exists a uniquely determined "error" function $\mu(z)$ holomorphic in $\mathbb{C}^-$ with the following properties*:

(a) $$\Gamma(z) = \sqrt{2\pi} z^{z-1/2} e^{-z} e^{\mu(z)} \quad \textit{for all } z \in \mathbb{C}^-.$$

(b) *For every δ, $0 < \delta \leq \pi$, and all $z = re^{i\varphi}$ satisfying $|\varphi| \leq \pi - \delta$ we have*:

$$|\mu(z)| \leq \frac{1}{8} \frac{1}{\sin^2 \frac{1}{2}\delta} \frac{1}{|z|}.$$

The uniqueness of $\mu(z)$ is trivial. Clearly we have

$$\mu(z) = \log \Gamma(z) - \left(z - \frac{1}{2}\right) \log z + z - \frac{1}{2} \log 2\pi,$$

where $\log z$ and $\log \Gamma(z)$ are defined by the integrals

$$\int_1^z \frac{d\zeta}{\zeta} \quad \text{and} \quad \int_1^z \frac{\Gamma'(\zeta)}{\Gamma(\zeta)} d\zeta, \; z \in \mathbb{C}^-.$$

The above equation for $\mu(z)$ does not reveal at all the *asymptotic* behaviour of $\mu(z)$. An elegant approach to $\mu(z)$ was given 1889 by T.-J. STIELTJES, [St]. One proceeds as follows. First we introduce the *real* functions

$$P(t) := t - [t] - \frac{1}{2} \quad \text{and} \quad Q(t) = \frac{1}{2}\left(t - [t] - (t - [t])^2\right),$$

where $[t]$ denotes the largest integer $\leq t$. The "sawtooth" function $P(t)$ is continuous in $\mathbb{R} \setminus \mathbb{Z}$. The function $Q(t)$ is continuous in $\mathbb{R}$; furthermore $-Q'(t) = P(t)$ in $\mathbb{R} \setminus \mathbb{Z}$ and $0 \leq Q(t) \leq \frac{1}{8}$ for all $t \in \mathbb{R}$. Both functions have period 1. Following STIELTJES we put

$$\mu(z) := -\int_0^\infty \frac{P(t)}{z + t} dt = \int_0^\infty \frac{Q(t)}{(z + t)^2} dt \quad \text{for } z \in \mathbb{C}^-. \tag{1}$$

Both integrals here converge locally uniformly in $\mathbb{C}^-$ to the same holomorphic function; this is seen as follows: Since

$$|z + t|^2 = (r + t)^2 \cos^2 \frac{1}{2}\varphi + (r - t)^2 \sin^2 \frac{1}{2}\varphi \quad \text{for } z = re^{i\varphi},$$

we have

$$|z + t| \geq (|z| + t) \cos \frac{1}{2}\varphi \geq (|z| + t) \sin \frac{1}{2}\delta, \quad \text{if } |\varphi| \leq \pi - \delta.$$

Therefore

$$\left|\frac{Q(t)}{(z+t)^2}\right| \le \frac{1}{8}\frac{1}{\sin^2\frac{1}{2}\delta}\frac{1}{(|z|+t)^2}.$$

Hence the second integral converges in $\mathbb{C}^-$ locally uniformly to a function $\mu(z)$ holomorphic in $\mathbb{C}^-$; furthermore (b) of the Theorem follows from integrating this inequality. Since

$$-\int_r^s \frac{P(t)}{z+t}dt = \frac{Q(t)}{z+t}\bigg|_r^s + \int_r^s \frac{Q(t)}{(z+t)^2}dt \quad \text{for } 0 < r < s < \infty,$$

the first integral also converges locally uniformly towards $\mu(z)$.

Using $P(t+1) = P(t)$ we obtain

$$\mu(z+1) = -\int_0^\infty \frac{P(t+1)}{z+t+1}dt = -\int_1^\infty \frac{P(t)}{z+t}dt = \mu(z) - \int_0^1 \frac{\frac{1}{2}-t}{z+t}dt.$$

The integrand on the right is the derivative of $(z+\frac{1}{2})\log(z+t) - t$. The function defined by (1) therefore satisfies the functional equation

$$\mu(z) - \mu(z+1) = \left(z + \frac{1}{2}\right)\log\left(1+\frac{1}{z}\right) - 1, \; z \in \mathbb{C}^-, \tag{2}$$

in which log denotes the principal branch of the logarithm in $\mathbb{C}^-$.

9. PROOF OF THE EQUATION $\Gamma(z) = \sqrt{2\pi}z^{z-1/2}e^{-z}e^{\mu(z)}$. We consider in $\mathbb{C}^-$ the holomorphic function $F(z) := z^{z-1/2}e^{-z}e^{\mu(z)}$. Using 8(2) we obtain

$$F(z+1) = (z+1)^{z+1/2}e^{-z-1}e^{\mu(z)-(z+1/2)\log(1+1/z)+1}.$$

Now $\log(1+1/z) = \log(1+z) - \log z$ for all $z \in \mathbb{C}^-$. Hence we get

$$F(z+1) = e^{-z}e^{\mu(z)+(z+1/2)\log z} = z^{z+1/2}e^{-z}e^{\mu(z)} = zF(z).$$

Furthermore, $F(z)$ is bounded in $S = \{z \in \mathbb{C}\colon 1 \le \operatorname{Re} z < 2\}$. Clearly $e^{\mu(z)}$ is bounded there by 8,b). For all $z = x + iy = |z|e^{i\varphi}, |\varphi| < \pi$, we may write

$$|z^{z-1/2}e^{-z}| \le |z|^{x-1/2}e^{-y\varphi}.$$

If $z \in S$ and $|y| \ge 2$, we have $x - \frac{1}{2} \le 2$ and $|z| \le 2|y|$, hence $|z^{z-1/2}| \le 4y^2$. Since $x < |y|$, we have $\tan|\varphi| = |y|/x > 1$, hence $|\varphi| > \frac{1}{4}\pi$ and consequently $e^{-y\varphi} \le e^{-\pi|y|/4}$. Therefore, $|z^{z-1/2}e^{-z}| \le 4y^2e^{-\pi|y|/4}$. Since this function is small for large $|y|$, the function $z^{z-1/2}e^{-z}$ is also bounded in S.

WIELANDT's theorem now yields $\Gamma(z) = az^{z-1/2}e^{-z}e^{\mu(z)}$ with a constant a. By LEGENDRE's formula 6(2) we find (after obvious cancellations)

$$\sqrt{2\pi e}\,e^{\mu(2z)-\mu(z)-\mu(z+1/2)} = a\left(1+\frac{1}{2z}\right)^z.$$

Since $\lim_{x\to\infty}\mu(x) = 0$ and $\lim_{x\to\infty}(1+1/2x)^x = \sqrt{e}$ we get $a = \sqrt{2\pi}$. q.e.d.

It should be noted that there are better estimates for $\mu(z)$ than the one obtained in Theorem 8.b), e.g. the factor $\frac{1}{8}$ can be replaced by $\frac{1}{12}$. For details we refer to $[\mathbf{R}_1]$.

REFERENCES

[A] Artin, E.: Einführung in die Theorie der Gammafunktion, Teubner 1931 [English translation by M. BUTLER: The Gamma Function, Holt, Rinehart and Winston 1964].

[Bi] Birkhoff, G. D.: Note on the gamma function, Bull. AMS 20, 1–10 (1914).

[BM] Bohr, H. and J. Mollerup: Laerebog i matematisk Analyse, vol. 3, Kopenhagen 1922.
[D] Dedekind, R.: Über die Elemente der Theorie der Eulerschen Integrale, Inauguraldissertation Göttingen 1852, Ges. Math. Werke, 1, 1–26, Vieweg Braunschweig 1930.
[G] Gauss, C. F.: Disquisitiones generales circa seriam infinitam ..., Werke 3, 123–162, edited by Königl. Ges. Wiss. Göttingen 1876.
[K] Knopp, K.: Funktionentheorie II, 5. edition, Sammlung Göschen 703, de Gruyter Berlin 1941 [English translation by F. BAGEMIHL: Theory of Functions, Dover 1947].
[R_1] Remmert, R.: Funktionentheorie II, 2. edition, Springer 1995.
[R_2] Remmert, R.: Wielandt's characterization of the Γ-function, in WIELANDT's *Mathematical Works*, vol. 2, de Gruyter (to appear).
[St] Stieltjes, T.-J.: Sur le developpement de log $\Gamma(a)$, Journ. Math. pur. appl. (4)5, 425–44 (1889); Œuvres 2, Noordhoff 1918, 211–230; 2. ed. Springer 1993, 215–234.

Mathematisches Institut der Universität
48149 Münster
Einsteinstraße 62
Germany

Referee's Report: This paper contains much that is new and much that is true. Unfortunately, that which is true is not new and that which is new is not true.

From Howard W. Eves, *Return to Mathematical Circles*, Prindle, Weber and Schmidt, Boston, 1988, p. 158.

"...the source of all great mathematics is the special case, the concrete example. It is frequent in mathematics that every instance of a concept of seemingly great generality is in essence the same as a small and concrete special case."

Paul Halmos, *I Want to be a Mathematician*, MAA Spectrum, Washington, D.C., 1985, p. 324.

Charting the 3-Sphere—An Exposition for Undergraduates

Louis Zulli

Most people are familiar with the use of latitude and longitude in locating and naming points on the surface of the Earth, which, mathematically at least, is considered to be a perfectly round two-dimensional sphere. This article is about a similar coordinate system on the 3-sphere S^3, the set of points in four-dimensional Euclidean space that lie exactly one unit from the origin. S^3 is a rich and beautiful space, and an exploration of it, even at the level of this introductory article, involves a wealth of interesting mathematics. Indeed, here are just some of the mathematical players who will appear in the brief exposition we're about to present: numbers—real, complex and quaternion; matrices—real and complex, orthogonal and unitary; linear algebra—vector spaces, subspaces, inner products, traces, eigenvalues, eigenvectors, diagonalization and the matrix exponential; group theory—conjugation and conjugacy classes; geometry—intrinsic distance, tangent spaces and the exponential map. Quite a lot of mathematics in a short expository paper about a single three-dimensional space—and we could have gone much further. In fact, the main difficulty in writing this paper was knowing when to say 'Enough!'.

Which brings me to the point of this paper. I wrote this paper as a paper to be studied by undergraduate Mathematics Majors in a Senior Seminar. Of course, I hope that this article will be read by other people in other contexts as well, but my focus while writing this paper was fixed on the Senior Seminar, and I think that this focus is reflected in the paper's style. In particular, this paper is meant to be read and discussed by a group of students, so that all might benefit from the sharing of knowledge, and so that confusions might be quickly resolved and obstacles overcome. (I also assumed that there would be a faculty member present to help the students navigate through some of the paper's more treacherous passages.) I did not attempt to make this paper either self-contained or linear. This is because few topics in mathematics are truly self-contained, and because mathematics as a whole is decidedly non-linear. Throughout the paper there are terms and facts borrowed from many branches of mathematics, and there is also the occasional tidbit, usually parenthesized, meant to entice the readers to explore new territories in the mathematical kingdom. There are also many verifications left to the readers; most of these verifications are calculations. Having said all this, I feel that I should add that I tried to organize the paper so as to assist the readers in learning the material, and that I attempted to explain, at least informally, most of the technical terms in the article. I also tried to give proofs, or sketches of proofs, of most of the claims made within. At the end of the paper is a short list of references, which I include mainly as an act of basic decency. My hope is that the readers already have their own favorite sources of mathematical knowledge, and that they will consult these favorite sources as needed. I was inspired to write this

article after reading the first dozen pages of **[1]**. In some sense, those pages provided an outline for this paper. (After writing this article I discovered a similar treatment in Chapter 8 of **[2]**. A reading of that chapter would be an excellent supplement to this exposition.)

1. S^1 **AND** $SO(2)$. Before confronting the 3-sphere S^3 in $\mathbf{R}^4$, let's first consider the familiar 1-sphere (unit circle) S^1 in $\mathbf{R}^2$. (This circle will reappear later, as the "great circle" in S^3 containing the "Prime Meridian.") Let $z = x + iy$ be a complex number, which we will identify with the point $z = (x, y)$ in $\mathbf{R}^2$. Sometimes we'll also think of $z = (x, y)$ as being the vector $\vec{z}$ in $\mathbf{R}^2$ from the origin to (x, y). There is a well-known correspondence (bijection) between the complex numbers and certain 2×2 matrices of real numbers:

$$z = x + iy \in \mathbf{C} = \mathbf{R}^2 \leftrightarrow Z = \begin{pmatrix} x & -y \\ y & x \end{pmatrix} \in M_2(\mathbf{R}).$$

(The notational convention employed here is used throughout this article. That is, points in spaces are denoted by lowercase letters and matrices by uppercase letters. The passage from a point to a matrix representation of that point is indicated by capitalization.) This correspondence is actually a field isomorphism, meaning that it respects addition and multiplication. Via this correspondence we obtain a matrix model of $\mathbf{C}$, which we can also view as a matrix model of $\mathbf{R}^2$.

In $\mathbf{R}^2$ there is an inner product, the ordinary dot product of vectors, and this inner product is quite helpful when one does geometry in $\mathbf{R}^2$. For vectors $\vec{z}_1 = (x_1, y_1)$ and $\vec{z}_2 = (x_2, y_2)$ in $\mathbf{R}^2$, $\vec{z}_1 \cdot \vec{z}_2 = \langle \vec{z}_1, \vec{z}_2 \rangle = x_1 x_2 + y_1 y_2$. If we treat these same vectors as complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, then the dot product can be computed as $\langle z_1, z_2 \rangle = \frac{1}{2}(z_1 \bar{z}_2 + \bar{z}_1 z_2)$. What is the formula for this standard inner product in the matrix model of $\mathbf{R}^2$? It is not difficult to verify that $\langle Z_1, Z_2 \rangle = \frac{1}{2}\mathrm{tr}(Z_1 Z_2^T)$ exactly corresponds to $\langle z_1, z_2 \rangle = \frac{1}{2}(z_1 \bar{z}_2 + \bar{z}_1 z_2)$, where Z_2^T denotes the transpose of Z_2, and $\mathrm{tr}(Z_1 Z_2^T)$ denotes the trace of $Z_1 Z_2^T$, the sum of the elements on the main diagonal of $Z_1 Z_2^T$. Using the inner product on $\mathbf{R}^2$ we can compute the norm (length) $|z|$ of a complex number (vector) $z \in \mathbf{R}^2$. By definition, $|z|^2 = \langle z, z \rangle$. In our matrix model of $\mathbf{R}^2$ this becomes $|Z|^2 = \frac{1}{2}\mathrm{tr}(ZZ^T)$, which simplifies to $\det(Z)$, the determinant of Z, because of the special form which Z has. So $|Z|^2 = \det(Z)$.

By definition, $S^1 = \{z \in \mathbf{C}\colon |z| = 1\} = \{\vec{z} \in \mathbf{R}^2\colon |\vec{z}| = 1\}$; this is the unit circle in the plane $\mathbf{R}^2$. In the matrix model of $\mathbf{R}^2$ this corresponds to $\{Z = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}\colon \det(Z) = 1\}$. This set of matrices can also be described as $\{M \in M_2(\mathbf{R})\colon MM^T = I, \det(M) = 1\}$. A square matrix M of real numbers that satisfies the equation $MM^T = I$ is called an *orthogonal matrix*, and its determinant is necessarily 1 or -1. Those orthogonal matrices with determinant 1 are called *special orthogonal matrices*. Thus $S^1 \subseteq \mathbf{C} = \mathbf{R}^2$ corresponds to the group of 2×2 special orthogonal matrices, which is denoted $SO(2)$. Furthermore, the correspondence is a group isomorphism, where the operation in both groups is multiplication. It is worth knowing that $SO(n)$ can also be thought of as the group of orientation-preserving linear isometries of $\mathbf{R}^n$, that is, as the group of orientation-preserving rigid motions of $\mathbf{R}^n$ that fix the origin. In $\mathbf{R}^2$ the only such motions are rotations about the origin, and there is one such rotation for each point on the unit circle S^1. So perhaps it's not surprising that S^1 is isomorphic to the group $SO(2)$.

2. S^3 AND $SU(2)$. Now let $v = x + iy + jz + kt$ be a quaternion, which we will identify with the point $v = (x, y, z, t)$ in $\mathbf{R}^4$, and sometimes with the vector $\vec{v} = (v_1, v_2) \in \mathbf{C}^2$, where $v_1 = x + iy$ and $v_2 = z + it$. As above, there is a correspondence between the quaternions and certain 2×2 matrices of complex numbers. In this case, the correspondence is:

$$v = x + iy + jz + kt \in \mathbf{H} = \mathbf{R}^4 \leftrightarrow V = \begin{pmatrix} v_1 & v_2 \\ -\bar{v}_2 & \bar{v}_1 \end{pmatrix} \in M_2(\mathbf{C}),$$

where $v_1 = x + iy$ and $v_2 = z + it$. This correspondence preserves addition and multiplication; it is an isomorphism of division rings. (A division ring is a "possibly non-commutative field"—multiplication might not be commutative but all the other field axioms are satisfied.) Via this correspondence we obtain a matrix model of $\mathbf{R}^4$. In this model, the ordinary Euclidean inner product becomes $\langle V_1, V_2 \rangle = \frac{1}{2}\mathrm{tr}(V_1\bar{V}_2^T)$, and, as above, $|V|^2 = \det(V)$.

By definition, $S^3 = \{v \in \mathbf{H}\colon |v| = 1\} = \{\vec{v} \in \mathbf{R}^4\colon |\vec{v}| = 1\}$; this is the unit 3-sphere in four-dimensional space. Note that S^3 is a (non-commutative) group via quaternion multiplication. In the matrix model of $\mathbf{R}^4$, S^3 corresponds to $\{V = \begin{pmatrix} v_1 & v_2 \\ -\bar{v}_2 & \bar{v}_1 \end{pmatrix} \in M_2(\mathbf{C})\colon \det(V) = 1\}$, which can also be described as $\{M \in M_2(\mathbf{C})\colon M\bar{M}^T = I, \det(M) = 1\}$. This matrix group is denoted $SU(2)$, and is called the two-dimensional *special unitary group.* The correspondence $S^3 \leftrightarrow SU(2)$ defined above is a group isomorphism. Because of this, we will often blur the distinction between S^3 and $SU(2)$. In particular, we will often consider points in S^3 to be matrices, labeling them with uppercase letters.

3. DISTANCE, PARALLELS AND MERIDIANS IN S^3. We can already compute the distance in $\mathbf{R}^4$ between points v and w in S^3. This is just $|v - w|$, or, in our matrix formulation, $[\det(V - W)]^{1/2}$. This distance is simply the length of a vector in $\mathbf{R}^4$ from v to w. But such a vector intersects S^3 only at v and w, and its length is not the distance in S^3 from v to w. (The distinction between these two notions of distance is easily seen in analogy: The distance across the surface of the Earth from Lisbon to Vladivostok is greater than the distance between those cities were tunneling through the Earth allowed.) How can we calculate $d_{S^3}(v, w)$, the so-called *intrinsic distance* from v to w in S^3? Consider a two-dimensional plane Π containing v, w and the origin O in $\mathbf{R}^4$. This plane will be unique unless $w = -v$, but uniqueness of Π is not required. What is $\Pi \cap S^3$, the intersection of Π and S^3 in $\mathbf{R}^4$? $\Pi \cap S^3$ lies within both Π and S^3, but it is easier to analyze within Π, where it is just the set of points in Π one unit from the origin O. Thus $\Pi \cap S^3$ is just the unit circle in Π. In S^3, the circle $\Pi \cap S^3$ is called a *great circle,* and the key fact that we need is that this circle provides the shortest path within S^3 from v to w. Accepting this as true, it is then easy to compute $d_{S^3}(v, w)$, by focusing once again on the plane Π. Since v and w lie on the unit circle in Π, the distance between them along that circle is exactly the radian measure θ of the (smaller) undirected angle vOw in Π, by the very definition of radian measure. See Figure 1. From the geometric dot product formula, we obtain $\theta = \cos^{-1}(\langle v, w \rangle)$, since v and w are unit vectors. Thus $d_{S^3}(v, w) = \cos^{-1}(\langle v, w \rangle)$, or, in our matrix formulation, $d_{S^3}(V, W) = \cos^{-1}(\frac{1}{2}\mathrm{tr}(V\bar{W}^T))$.

With this distance formula in hand, we can begin charting S^3. Since the identity matrix I corresponds to the point $(1, 0, 0, 0) \in S^3 \subseteq \mathbf{R}^4$, the positive x-axis in $\mathbf{R}^4$

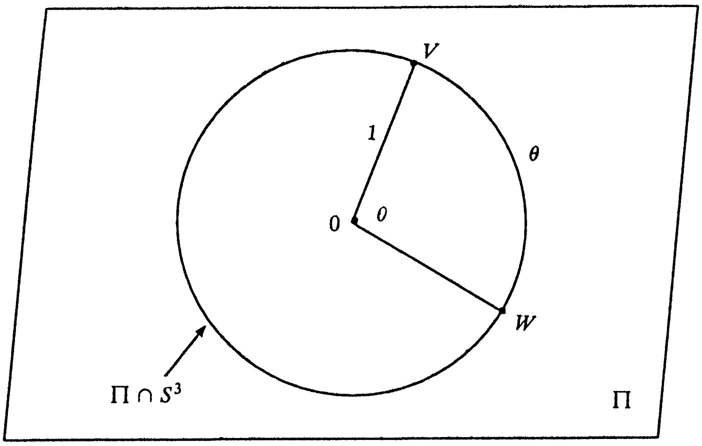


Figure 1

passes through I. We will view I as the "North Pole" in our model of S^3—not surprisingly, $-I$ will be the "South Pole." For $L \in [0, \pi]$, let $\mathscr{P}_L = \{V \in S^3: d_{S^3}(V, I) = L\}$. We will call $\mathscr{P}_L$ the *parallel in* S^3 *at latitude* L. Referring to the distance L as latitude is quite reasonable, since not only is L the distance in S^3 from V to I, but it's also the radian measure of angle IOV. (So we'll be measuring our latitudes down from the North Pole, rather than up and down from the Equator as is customary on Earth.) $\mathscr{P}_0 = \{I\}$, that is, I is the only point in S^3 at latitude 0, since it's the only point in S^3 whose distance from I is zero. $\mathscr{P}_\pi = \{-I\}$, although perhaps it is not immediately apparent that $-I$ is the only point in S^3 at distance π from I. We will refer to $\mathscr{P}_{\pi/2}$ as the *Equator* in S^3, and use the symbol $\mathscr{E}$ to denote this special parallel. It is helpful to have several descriptions of $\mathscr{P}_L$ at one's fingertips, in particular: $\mathscr{P}_L = \{V \in S^3: \frac{1}{2}\mathrm{tr}(V) = \cos(L)\}$ and $\mathscr{P}_L = \{(x, y, z, t) \in S^3 \subseteq \mathbf{R}^4: x = \cos(L)\}$. From the first of these two reformulations we conclude that all points on a given parallel in S^3 have the same trace, and that V lies on the Equator $\mathscr{E}$ in S^3 if and only if $\mathrm{tr}(V) = 0$. From the second reformulation we see that $\mathscr{P}_L$ is a two-dimensional sphere of radius $\sin(L)$ centered at $(\cos(L), 0, 0, 0)$ within the hyperplane $x = \cos(L)$ in $\mathbf{R}^4$. (If $L = 0$ or $L = \pi$ this sphere has radius 0, so $\mathscr{P}_0$ and $\mathscr{P}_\pi$ are indeed one point sets.) In particular, $\mathscr{E}$ is just the unit 2-sphere in the three-dimensional subspace of $\mathbf{R}^4$ spanned by the y, z and t axes—or equivalently, $\mathscr{E}$ is the set of pure imaginary unit quaternions, that is, the unit quaternions of the form $iy + jz + kt$. From the second reformulation of $\mathscr{P}_L$ we can also conclude that $S^3 = \cup_{L \in [0, \pi]} \mathscr{P}_L$, that is, that every point in S^3 lies on some parallel. This is because $x \in [-1, 1]$ for every point $(x, y, z, t) \in S^3 \subseteq \mathbf{R}^4$, so $x = \cos(L)$ for some $L \in [0, \pi]$. Clearly $\mathscr{P}_L \cap \mathscr{P}_{L'} = \varnothing$ if $L \neq L'$, so in fact each point in S^3 lies on a unique parallel.

Now let $E \in \mathscr{E}$. By the *meridian* $\mathscr{M}_E$ *at longitude* E we mean the set of points in S^3 of the form $\cos(\theta)I + \sin(\theta)E$, for $\theta \in [0, \pi]$. (Notice that, once again, we're deviating slightly from the convention used on Earth. To us, the longitude of a meridian in S^3 is simply the point of intersection of that meridian with the Equator. If an analogous system were used on Earth, then the meridian containing Buffalo, N.Y. would be called "Quito, Ecuador" rather than "78°W.") It is easy to check that $\mathscr{M}_E$ intersects $\mathscr{P}_L$ only at the point $\cos(L)I + \sin(L)E$, and that

$\mathscr{M}_E \cap \mathscr{M}_{E'} = \{\pm I\}$ if $E \neq E'$. To see that $S^3 = \cup_{E \in \mathscr{E}} \mathscr{M}_E$, that is, that every point $V \in S^3$ lies on at least one meridian, one could argue geometrically as follows: Choose a plane Π in $\mathbf{R}^4$ containing I, V and the origin O. As before, Π will be unique if $V \neq \pm I$, but uniqueness of Π is not required. Let $\mathbf{R}^3$ denote the three-dimensional subspace of $\mathbf{R}^4$ spanned by the y, z and t axes, so that $\mathbf{R}^3 \cap S^3 = \mathscr{E}$. Clearly $\Pi \not\subseteq \mathbf{R}^3$ since $I \in \Pi$ but $I \notin \mathbf{R}^3$. Since $\Pi \cap \mathbf{R}^3$ is necessarily a subspace of Π, either $\Pi \cap \mathbf{R}^3 = \{O\}$ or $\Pi \cap \mathbf{R}^3$ is a line through O. The first of these possibilities is eliminated by dimensional considerations, so $\Pi \cap \mathbf{R}^3$ is a line in $\mathbf{R}^3$ containing O. This line intersects $\mathscr{E}$, the unit sphere in $\mathbf{R}^3$, in two points, E and $-E$. Since $(\Pi \cap \mathbf{R}^3) \cap \mathscr{E} = (\Pi \cap S^3) \cap \mathscr{E}$, the great circle $\Pi \cap S^3$, which contains V, intersects $\mathscr{E}$ at $\pm E$. Since I and E are orthogonal points on this circle in Π, each point on the circle can be expressed as $\cos(\theta)I + \sin(\theta)E$, for some $\theta \in [0, 2\pi]$. In particular, $V = \cos(\theta)I + \sin(\theta)E$ for some $\theta \in [0, 2\pi]$. If $\theta \in [0, \pi]$ then $V \in \mathscr{M}_E$, and we're done. Otherwise, $\theta = 2\pi - \theta'$ with $\theta' \in [0, \pi]$, and substituting for θ, we obtain $V = \cos(\theta')I + \sin(\theta')(-E)$, so $V \in \mathscr{M}_{-E}$.

Although this geometric argument demonstrates that each $V \in S^3$ lies on a meridian, it may not provide the best procedure for actually calculating the point E (or $-E$) naming that meridian. In the next section we'll see another way to find the meridian in question. For the moment though, we have established that each point $V \in S^3 - \{\pm I\}$ lies on a unique parallel $\mathscr{P}_L$ and a unique meridian $\mathscr{M}_E$, and that $\mathscr{M}_E \cap \mathscr{P}_L = \{V\}$. Thus each such V has a unique latitude $L \in [0, \pi]$ and longitude $E \in \mathscr{E}$, and furthermore, no other point in S^3 has the same latitude and longitude.

4. CONJUGATION IN S^3. S^3 is a much more interesting space than either S^1 or S^2, partly because it's of a higher dimension than either of those spaces, but mostly because of its group structure, which as we shall see, meshes quite nicely with the geometry of the space. (As we saw in Section 1, the circle S^1 is also a group, but its multiplication is commutative, making the overall structure of S^1 somewhat less interesting than that of S^3. And, perhaps surprisingly, S^2 does not admit *any* group structure analogous to the structures on S^1 and S^3.) In particular, we need to understand how conjugation in S^3 interacts with the coordinate system we constructed in the previous section. To this end, let $C \in S^3$, and let $C_*: S^3 \to S^3$ denote the map $V \mapsto CVC^{-1}$, which is called *conjugation by* C. Because $C \in S^3, C\overline{C}^T = I$, so C_* is also the map $V \mapsto CV\overline{C}^T$. For each $C \in S^3, C_*$ is a group isomorphism—it's easily seen to be a homomorphism and its inverse is $(C^{-1})_*$. But, even better, C_* is an *isometry* of S^3, it preserves the distance between points in S^3. That is, for any V and W in S^3, $d_{S^3}(C_*(V), C_*(W)) = d_{S^3}(V, W)$. (This easily verified equality is a consequence of the fact that conjugation does not change the trace of a matrix.) Because conjugation is an isometry and $\mathscr{P}_L = \{V \in S^3\colon\ d_{S^3}(I, V) = L\}, C_*$ maps $\mathscr{P}_L$ isometrically onto itself for each $L \in [0, \pi]$ and each $C \in S^3$. In fact, the collection of parallels in S^3 is precisely the set of conjugacy classes in S^3. That is, V and W are conjugate in S^3 (meaning $W = C_*(V)$ for some $C \in S^3$) if and only if V and W lie in the same parallel, that is, if and only if V and W have the same latitude in S^3. Before establishing this fact, it is helpful to expand our terminology and notation. For $L \in [0, \pi]$, let D_L denote the point $\begin{pmatrix} e^{iL} & 0 \\ 0 & e^{-iL} \end{pmatrix} \in S^3$, and let $\mathscr{D} = \{D_L\colon\ L \in [0, \pi]\}$. Then $\mathscr{D}$ is also the meridian $\mathscr{M}_{D_{\pi/2}}$, and we'll refer to $\mathscr{D}$ as the *Prime Meridian* in S^3. (We use the letter '$\mathscr{D}$' because all the points D_L in $\mathscr{D}$ are diagonal matrices.) Since $D_{\pi/2}$ corresponds to $(0, 1, 0, 0)$ in $\mathbf{R}^4$, the positive y-axis passes through this point, and

thus $\mathscr{D}$ is contained in the unit circle $S^1 \subseteq \mathbf{R}^2$. In fact, $\mathscr{D}$ is precisely the unit semi-circle in the closed upper half-plane of $\mathbf{R}^2$. It is worth noting that, for each $L \in [0, \pi]$, $\mathscr{D}$ intersects the parallel $\mathscr{P}_L$ at the single point D_L. In particular, the latitude of D_L is L.

We claim that $\mathscr{P}_L$ is precisely the conjugacy class of D_L in S^3. We already know that every conjugate of D_L lies in $\mathscr{P}_L$. What we need to show is that each point V in $\mathscr{P}_L$ is conjugate to the diagonal matrix D_L. This can be done as follows: First off, the result is clearly true for $L = 0$ and $L = \pi$, so we'll assume that $L \in (0, \pi)$. Since V is a 2×2 matrix, the characteristic polynomial of V is $\lambda^2 - \operatorname{tr}(V)\lambda + \det(V)$. Because $V \in \mathscr{P}_L \subseteq S^3$, this becomes $\lambda^2 - (2\cos(L))\lambda + 1$, and the quadratic formula yields two distinct eigenvalues, $\lambda_1 = \cos(L) + i\sin(L) = e^{iL}$ and $\lambda_2 = \bar{\lambda}_1 = \cos(L) - i\sin(L) = e^{-iL}$. Equip the complex vector space $\mathbf{C}^2$ with its usual hermitian inner product, and choose a unit eigenvector (z_1, z_2) for V corresponding to the eigenvalue λ_1. Let $C = \begin{pmatrix} z_1 & -\bar{z}_2 \\ z_2 & \bar{z}_1 \end{pmatrix}$. Then $\det(C) = 1$ because (z_1, z_2) is a unit vector in $\mathbf{C}^2$, so $C \in S^3$. Since

$$(C^{-1}VC)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = (C^{-1}V)\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = (\lambda_1 C^{-1})\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \lambda_1\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ 0 \end{pmatrix},$$

$C^{-1}VC = \begin{pmatrix} \lambda_1 & \alpha \\ 0 & \beta \end{pmatrix}$, for some $\alpha, \beta \in \mathbf{C}$. But $C^{-1}VC \in S^3$, so we must have $\alpha = 0$ and $\beta = \bar{\lambda}_1$. That is, $C^{-1}VC = D_L$, or equivalently, $V = CD_LC^{-1}$, establishing the claim. (It is worth remarking that the point C in S^3 conjugating D_L to V is not unique. In fact, the set of points in S^3 which conjugate D_L to V is precisely CS^1, the coset containing C of the subgroup $S^1 \subseteq S^3$. Continued exploration along these lines leads to alternative measures of longitude in S^3, and to the intriguing equation $\mathscr{E} = S^2 = S^3/S^1$.)

Having just examined the effect of conjugation on parallels in S^3, we now turn our attention to meridians. Here there is less work to do. The basic fact is: For each $C \in S^3$ and $E \in \mathscr{E}$, $C_*(\mathscr{M}_E) = \mathscr{M}_{C_*(E)}$. That is, conjugation by C carries the meridian at longitude E isometrically onto the meridian at longitude $C_*(E)$. Thus, for each $C \in S^3$, C_* can be thought of as a "rigid rotation" of S^3 about its polar axis. See Figure 2 for a schematic picture. Putting things together, what we have

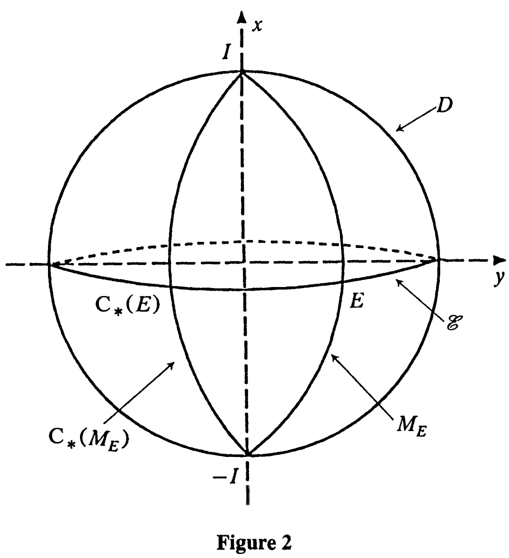


Figure 2

shown is that, for each $C \in S^3$, conjugation by C carries the point at latitude-longitude (L, E) to the point at latitude-longitude $(L, C_*(E))$.

Using conjugation we can now outline a second approach to finding the longitude of an arbitrary point $V \in S^3 - \{\pm I\}$. Let L denote the latitude of V. As was shown above, there is a C in S^3 for which $C_*(D_L) = V$. Let $E = C_*(D_{\pi/2})$. Then E has the same longitude as V, since D_L has the same longitude as $D_{\pi/2}$ and conjugation takes meridians to meridians. Furthermore, $E \in \mathscr{E}$, since $D_{\pi/2} \in \mathscr{E}$ and conjugation preserves latitudes. Thus the longitude of V is E. For an explicit formula, let $(z_1, z_2) \in \mathbf{C}^2$ be a unit eigenvector for V with eigenvalue e^{iL}. Then

$$E = \begin{pmatrix} i(z_1\bar{z}_1 - z_2\bar{z}_2) & 2iz_1\bar{z}_2 \\ 2i\bar{z}_1 z_2 & -i(z_1\bar{z}_1 - z_2\bar{z}_2) \end{pmatrix}.$$

5. THE EXPONENTIAL MAP. Let us return for a moment to the circle $S^1 \subseteq \mathbf{C} = \mathbf{R}^2$. By the *tangent space to S^1 at* 1, denoted $T_1(S^1)$, we mean the set of all vectors in $\mathbf{R}^2$ that are tangent to S^1 at 1. Although it is illustrative to draw vectors in $T_1(S^1)$ as tangent to S^1 at 1, we will consider tangent vectors as actually emanating from the origin in $\mathbf{R}^2$. Thus each vector in $T_1(S^1)$ corresponds to a vector lying in the y-axis, or, in the language of complex numbers, to an imaginary number iy. See Figure 3. Via this identification, $T_1(S^1)$ becomes a one-dimensional subspace of $\mathbf{R}^2$, and we can use the identification to define a map from $T_1(S^1)$ onto S^1. Namely, given iy representing a tangent vector to S^1 at 1, map iy to $e^{iy} \in S^1$. This aptly named *exponential map* $T_1(S^1) \to S^1$ takes the zero vector in $T_1(S^1)$ to $1 \in S^1$, takes the upward pointing tangent vector of length $\pi/2$ to $i \in S^1$, and takes the downward pointing tangent vector of length $\pi/2$ to $-i \in S^1$. Both tangent vectors of length π in $T_1(S^1)$ are mapped to $-1 \in S^1$. Putting it somewhat informally, the exponential map "wraps the tangent space $T_1(S_1)$ around S^1." In particular, it maps the interval $[-\pi i, \pi i] \subseteq T_1(S^1)$ onto S^1, and the mapping is one-to-one when restricted to $(-\pi i, \pi i)$. This map $T_1(S^1) \to S^1$ is a very special case of a

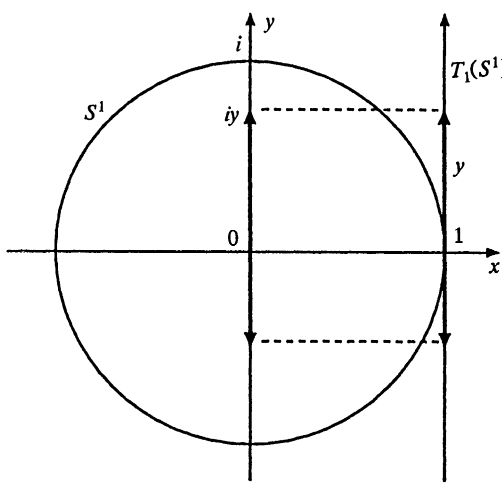


Figure 3

standard construction in differential geometry. In fact, it is possible to define an exponential map $T_p(M) \to M$ at any point p on any smooth manifold M. (Try envisioning what the exponential map would look like on the 2-sphere and on the torus.)

Now let's return to S^3. Since S^3 is a three-dimensional manifold, $T_I(S^3)$ is isomorphic to $\mathbf{R}^3$. That is, there are three linearly independent directions for tangent vectors to S^3 at I. As before we can (try to) picture vectors in $T_I(S^3)$ as being tangent to S^3 at I, but in actuality we'll treat these vectors as emanating from the origin O in $\mathbf{R}^4$. Since vectors tangent to S^3 at I are necessarily orthogonal to the vector from O to I, we can think of $T_I(S^3)$ as the collection of all vectors in $\mathbf{R}^4$ perpendicular to the x-axis, since this is the axis containing $I = (1,0,0,0)$. Thus $T_I(S^3)$ can be identified with the three-dimensional subspace of $\mathbf{R}^4$ spanned by the y, z and t axes. In the language of quaternions, $T_I(S^3)$ corresponds to the pure imaginary quaternions, those quaternions of the form $iy + jz + kt$. This model allows a simple identification of $T_I(S^3)$ with our familiar $\mathbf{R}^3$—simply view the unit quaternions i, j and k as the standard unit vectors $\vec{i}$, $\vec{j}$ and $\vec{k}$ in $\mathbf{R}^3$. (Via this identification with $\mathbf{R}^3$, we see that $T_I(S^3)$ is more than just a real three-dimensional vector space, it's a "Lie algebra" over $\mathbf{R}$, where the "bracket product" of tangent vectors is just the classical cross product in $\mathbf{R}^3$.) In our matrix model of $\mathbf{R}^4$, $T_I(S^3) = \{\begin{pmatrix} iy & v_2 \\ -\bar{v}_2 & -iy \end{pmatrix}: y \in \mathbf{R},\, v_2 \in \mathbf{C}\}$, where we no longer require that the matrices have determinant 1, since tangent vectors need not be unit vectors.

Using this matrix formulation of $T_I(S^3)$, it is easy to define an exponential map $T_I(S^3) \to S^3$. It's just the map $V \mapsto \exp(V)$, where $\exp(V)$ denotes the matrix exponential of V, the same matrix exponential that is commonly introduced in linear algebra to study systems of linear differential equations. In fact, for a tangent vector $V \neq 0$ in the closed ball of radius π about the origin in $T_I(S^3)$, we claim that $\exp(V)$ is the point in S^3 at latitude $|V|$ and longitude $V/|V| \in \mathscr{E}$. (Of course, if $V = 0$ then $\exp(V) = I$.) To verify this claim, let V be such a tangent vector, and let E denote the point $V/|V| \in \mathscr{E}$. Then $\exp(V) = \exp(|V|E) = \exp(|V|CD_{\pi/2}C^{-1})$ (since all points in $\mathscr{E}$ are conjugate to $D_{\pi/2}$) $= \exp(C(|V|D_{\pi/2})C^{-1}) = C\exp(|V|D_{\pi/2})C^{-1} = CD_{|V|}C^{-1}$, where the final two equalities reflect basic properties of the matrix exponential. All that remains is to identify the latitude and longitude of $CD_{|V|}C^{-1}$, which, for notational convenience, we'll denote X. But this is easy. Since the latitude of $D_{|V|}$ is $|V|$, so is the latitude of its conjugate X. And since C_* carries $D_{\pi/2}$ to E, it must carry $D_{|V|}$ to a point in $\mathscr{M}_E$, so that the longitude of X is E. This establishes the claim made above. What it tells us is that the matrix exponential $T_I(S^3) \to S^3$ maps the closed ball of radius π about the origin in $T_I(S^3)$ onto S^3, and that the restriction of this map to the interior of that ball is one-to-one. Indeed, radii of the ball are carried isometrically onto meridians in S^3, while concentric spheres (centered at the origin) within the ball are mapped diffeomorphically onto the parallels in S^3. So, in some sense, the exponential map $T_I(S^3) \to S^3$ "wraps $T_I(S^3)$ around S^3," just as the map $iy \mapsto e^{iy}$ did in the one-dimensional case. (In fact, the exponential map $T_I(S^3) \to S^3$ is an extension of the exponential map $T_I(S^1) \to S^1$.) The inverse of the exponential map carries $S^3 - \{-I\}$ diffeomorphically onto the open ball of radius π in $T_I(S^3)$. Thus it maps $S^3 - \{-I\}$, which lies in four-dimensional space, onto an open ball in $\mathbf{R}^3$. This is analogous to mapping the surface of the Earth, which lies in three-dimensional space, onto the two-dimensional page of an atlas, via equidistant polar projection. So the exponential map provides another approach to "charting" the 3-sphere.

6. CONCLUDING REMARKS. As I remarked in the introduction, there's still much more we could do. Discussions about Lie groups, group actions and fiber bundles follow naturally from the material presented in this paper, as does a more in-depth treatment of the geometry of S^3. But for this exposition, I think that now is the right time to say 'Enough!'. I hope that this article has provided you with a better understanding (and a usable model) of the 3-sphere, and that you are inspired to pursue further some of the ideas introduced above.

REFERENCES

1. S. Akbulut and J. McCarthy, *Casson's Invariant for Oriented Homology 3-Spheres An Exposition*, Mathematical Notes 36, Princeton University Press, Princeton, 1990.
2. M. Artin, *Algebra*, Prentice Hall, Englewood Cliffs, New Jersey, 1991.
3. C. Chevalley, *Theory of Lie Groups, I*, Princeton University Press, Princeton, 1946.
4. A. Knapp, *Lie Groups, Lie Algebras, and Cohomology*, Mathematical Notes 34, Princeton University Press, Princeton, 1988.
5. S. MacLane and G. Birkhoff, *Algebra*, 2nd ed., Collier Macmillan, New York, 1979.

Department of Mathematics
Rice University
P.O. Box 1892
Houston, TX 77251
zulli@math.rice.edu

A Ph.D. dissertation is a paper of the professor written under aggravating circumstances.

Attributed by G. Pòlya to Adolph Hurwitz, as quoted in Howard W. Eves, *Mathematical Circles Revisited*, Prindle, Weber and Schmidt, Boston, 1971.

Answer to Picture Puzzle

(p. 213)

Raoul Bott

Calculus: Reformed or Deformed?

Margaret Kleinfeld

Including more applications and more examples from the "real world" is one of the ideas bundled with calculus reform. I believe this is a move in the wrong direction. We should do *some* applications and examples, but we should do *fewer*, not more.

Applications involve other disciplines. Don't let them browbeat us into teaching every subject but our own! We have calculus for engineers, calculus for biological sciences, calculus for business—why can't we have calculus for mathematicians?

As for the "real world," mathematics is part of it. The real number system exists; studying it is a legitimate activity. Similarly for vector spaces or rings of matrices or.... .

The power of mathematics is abstraction. It was a great advance when prehistoric people realized they could think of numbers without using cows or stones. Now everyone insists we put the cows and stones back. The next advance was to realize we can do mathematics without numbers. Pure reasoning with symbols in place of arbitrary numbers lets one prove things that are true independent of a particular choice of numbers.

This second advance is the idea my students in linear algebra resist. "Do an example with numbers," they demand. They should already have progressed to this level of abstraction in high school algebra. Since they have not, we must struggle to help them reach it.

Learning by doing has limits. In my linear algebra course last year a student did badly on the first exam. "I don't understand it," he said. "I worked every problem in the book." I had told the class to work only what I assigned and to study carefully the review sheet. The next day he came back and said, "Boy, I should have paid more attention to that review sheet. Those exam questions were easy if you understood the theorems." Exactly! Happily, this student went on to do well in the class, but look at the bad study habits he had learned. Students have learned to tune out the teacher, and just work the problems.

I don't believe we have to motivate and justify everything we teach. For example, one year I was teaching a precalculus course, mainly aimed at liberal arts majors in order to fulfill a requirement. I had just proved that the square root of 2 is irrational, and I was in the midst of proving that the quotient of two integers always has a repeating or terminating decimal expansion (quite accessible since it only requires the division algorithm). One of my students raised his hand and asked, "what is this good for?" I let my annoyance show a little and replied: "If you just want to make change, you don't need it, but if you want to understand mathematics, the real number system is fundamental."

In this same class, I asked, "Which of the following are rational numbers?" One of the choices was 25/17. I expected this to be a gift, since the definition of a rational number was the quotient of two integers. One student, however, used his

calculator, which gave 1.4705882; he concluded the number was irrational because the decimal expansion was not repeating. If he had understood my theorem, he would have known that it must repeat. If he had followed my proof, he would have realized that the repeating segment could be up to sixteen digits long (and is in this case). Calculation and technology need to be accompanied by understanding.

When calculus reform speaks of increasing the teaching of concepts and decreasing the teaching of computational skills and cookbook methods, it goes in the right direction. But one cannot teach the *concept* of limit without using the epsilon-delta definition. Teaching such ideas intuitively does not make it easier for the student—it makes it harder to understand. Bertrand Russell has called the rigorous definition of limit and convergence the greatest achievement of the human intellect in 2000 years! The Greeks were puzzled by paradoxes involving motion; now they all become clear, because we have complete understanding of limits and convergence. Without the proper definition, things are difficult. With the definition, they are simple and clear.

A goal of calculus reform is to reach more students so that they can become scientifically and mathematically literate. An excellent goal. To achieve the goal, however, we need continually to distill and refine. We need to decide what we want students to know, and we need to develop it for them in the simplest, quickest way. Once long division was so difficult that only sages could do it. Then the positional notation was invented, and now school children do it with ease. Once Hamilton wrote volumes on quaternions. Now they are well understood and their properties can be developed in one lecture. Newton's *Principia* was extremely difficult, understood by only a few, and many things in the book were not rigorously defined or proved. Now in a slim volume the definitions and theorems can be made accessible. All one needs is a good understanding of the least upper bound axiom and the definition of limit, and the entire subject can be rigorously developed. This is how we help our students: by simplifying, improving, and finding new and better proofs. We don't need to pad this with lots of applications, examples from "real life," and motivation. We want to be lean, all right, but not lean and mean. We want lean and friendly calculus.

The calculus reform movement has said we want to replace the "Sage on the Stage" by the "Guide on the Side." Most people interpret this to mean less lecturing by the professor and more problem solving and projects by the students. Here again, it's the wrong direction. There is nothing wrong with students studying more, but a guide also can explain, and that is what a good lecture does. A teacher doesn't have to have the attitude, "See how smart I am? I can do these difficult problems and you can't!" Rather, the lecturer can communicate the attitude, "See these hard problems? I can do them, not because I'm so smart, but because brilliant people before me have figured out wonderful tricks which I have learned and which you can learn too!" The message should be, "If I can do it, you can do it too!" It's nice to let the students have the thrill of discovering some tricks on their own, but discovering all the tricks themselves discards our heritage.

We want our students to go beyond us. They can get off to a faster start than we did because we can pass on the knowledge from the past. Each idea and each proof presented in the most simple and elegant way we have ever seen it done! We should always have in mind the possibility that in each class there may be students who will go far beyond us.

The "new math" tried to make arithmetic more conceptual, to prepare students for more abstract thinking, for real mathematics. Calculus reform should do the same, but in many ways it seems to be going in the opposite direction.

Is the teacher a guide? Suppose you are in a foreign country, taking a tour of a cathedral, which is the pride and joy of the local populace. The guide stops by a beautiful stained glass window. He is explaining the pattern, the colors, how it was made. One of our group interrupts to ask, "What is this good for?" The guide feels angry and insulted. The question is rude. Another tourist pipes up, "I don't like cathedrals. Can't you show us the shopping center next door?" The whole group agrees they want to see the shopping center, so the guide takes them. Another group comes in. Again, they are soon bored and ask whether they can't tour the wine cellar instead. Again, the guide agrees to take them. Yet another group arrives, and once again doesn't want to see cathedrals. They want to see the steel mills. Finally, the guide refuses. "I am a cathedral guide and I love cathedrals," he says. "If you want to see steel mills, you will need another guide." The tourists all leave and the guide starts to walk away, when he notices that one young tourist has stayed behind. "This cathedral is the most beautiful thing I have ever seen," says the young tourist. "Would you please show me some more?"

Department of Mathematics
University of Iowa
14 MacLean Hall
Iowa City, IA 52242
mkleinf@math.uiowa.edu

"I was taught mathematics with a misplaced rigor, under the influence of the great mathematician Giuseppe Peano, but without adequate practical exercises. As my studies progressed, however, the teachers became of better quality."

Emilio Segrè's comment on his early mathematics education is in his autobiography, *A Mind Always in Motion*, University of California Press, Berkeley, 1993. Segrè also shares with his readers the following view of his later studies:

"The mathematics professor drove me crazy with Dedekind cuts. I learned rigorous proofs of seemingly obvious things, useless at my level, and at a time when with a little effort, I could have learned calculus, which would have been invaluable to me... I regret the effort spent in those years, so important for learning, on non-Euclidean geometry, number theory, and other subjects, completely omitting applied mathematics."

Submitted by Domenico Rosa, Teikyo Post University

The Method of Undetermined Generalization and Specialization

Illustrated with Fred Galvin's Amazing Proof of the Dinitz Conjecture

Doron Zeilberger

At the very beginning of our waning century, in what turned out to be the most influential mathematical address ever delivered, David Hilbert [H] said:

> *If we do not succeed in solving a mathematical problem, the reason frequently consists in our failure to recognize the more general standpoint from which the problem before us appears only as a single link in a chain of related problems.*

One paragraph later, he also said:

> *In dealing with mathematical problems, specialization plays, as I believe, a still more important part than generalization.*

Alas, all this is easier said than done. How does one find the 'right' generalization and specialization?

The answer is: just go ahead and start proving the conjecture. At first, leave the exact form of the generalization and/or specialization blank, and as you go along, see what kind of generalization/specialization would be required to make the proof work out. Keep 'guessing and erasing' until you get it done, just like doing a crossword puzzle.

I will illustrate this proof strategy using Fred Galvin's [G] recent brilliant proof of the Dinitz conjecture. Following a tradition that goes back to Euclid, Galvin presented his proof as a marvelous but 'static' completed edifice, just like the solution to yesterday's (or last Sunday's) puzzle, that hides all the trials and tribulations by which it was arrived. Not very useful for solving today's puzzle...

The Dinitz conjecture asserts that given n^2 arbitrary sets $A_{i,j}$ $(1 \leq i, j \leq n)$, each having n elements, then it is always possible to pick elements $a_{i,j} \in A_{i,j}$ such that $(a_{i,j})$ is a 'generalized Latin square', which means that each row and each column must have all its n entries distinct.

In other words, given a party of n boys and n girls, in which every boy must dance once with every girl, and such that each possible couple (i, j) knows how to dance (with each other[1]) only n dances, then out of the n^{n^2} ways of assigning dances to couples, there is at least one way in which each of the $2n$ individuals dances a different dance in each of his or her n performances.

[1] These are couples' dances and the two dancers should be able to coordinate their steps, so it is possible for Abe to be able to dance the tango with Alice but not with Barbara, although Barbara may be able to dance it with other boys.

Like many people, I first heard about the Dinitz conjecture **[ERT]** when Jeannette Janssen **[J]** brilliantly applied the powerful algebraic method of Alon and Tarsi **[AT]** to 'almost' prove it: she proved the analogous statement for 'non-square' rectangles.

As soon as I found out about the Dinitz conjecture, I was struck by its simplicity. Like so many times before, it seemed to me that there ought to be a 'simple' proof to such a simple statement, and I spent many hours trying, in vain, to prove it.

The reason I found my inability to prove the Dinitz conjecture so frustrating is that it appears to be 'intuitively obvious'. When all the sets $A_{i,j}$ are (pairwise) disjoint, then the statement is obvious. In the other extreme, when all the sets $A_{i,j}$ coincide, then it is also obvious: we have the problem of constructing an ordinary $n \times n$ Latin square. This can be constructed by looking at the multiplication table of any group of order n, in particular, the additive group of $\{0, 1, \dots, n-1\}$mod n:

$$\begin{matrix} 0 & 1 & \cdots & n-2 & n-1 \\ 1 & 2 & \cdots & n-1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ n-1 & 0 & \cdots & n-3 & n-2 \end{matrix} . \tag{L}$$

It is intuitively obvious that as there is less overlap between the sets $A_{i,j}$, there would have to be more options, which should make it even easier to construct a generalized Latin square.

The need to prove intuitively obvious statements is very common in analysis, which is why I find it such a forbidding subject. Indeed, this was one of the reasons why, shortly after my Ph.D., I decided to change fields from analysis to combinatorics, which I found much more gratifying, as there the gap between 'convincing yourself' and 'convincing the referee' is usually so much smaller.

Last spring, while I was still spending an hour a day trying to prove the Dinitz conjecture, I got an E-mail message from Herb Wilf, who forwarded an E-mail message from Richard Ehrenborg, who forwarded an E-mail message from Jeannette Janssen, who forwarded an E-mail message from Gil Kalai **[K]**[2] which contained a lucid and concise two-and-a-half page outline of Fred Galvin's proof, that Kalai had compiled.

When I finished reading and digesting the proof, I kicked myself. I felt that I could have found it myself, had I only followed Hilbert's advice, coupled with the 'crossword methodology' alluded to above. With the very generous help of Lady Hindsight, I will now describe how I (and you!) could have, and should have, found the very same proof, *without any prior knowledge of combinatorics or graph theory*.

I hope this presentation will enable you (or, better still, me!), to give an elementary proof (not using 'the class number formula for the Selmer group associated to the symmetric square representation of a modular lifting') of Wiles' theorem, and an elementary (and 'the first') proof of the Riemann Hypothesis.[3]

[2] I am pleased to be in the same connected component as Kalai, but I wish that Gil would draw a directed edge between him and myself, especially since there already exists an edge from me to him.

[3] RH is (almost) equivalent to the following elementary statement: Let a_n be the difference between the number of square-free integers between 1 and n with an even number of prime factors and the number of those with an odd number of prime factors, then for some constant A, $a_n \leq An^{9999/10000}$. This would already make you rich and famous. The full RH is equivalent to replacing the 9999/10000 by any number larger than 1/2.

THE METHOD OF UNDETERMINED COEFFICIENTS. Every mathematician (and electrical engineer) knows that in order to find a particular solution of a linear differential equation such as

$$y'' + y = x^2,$$

one writes $y = Ax^2 + Bx + C$, for some *undetermined* constants A, B, C. We don't know yet what they are, but we hope that they exist, and are constants. Assuming this, we plug it into the equation, getting

$$Ax^2 + Bx + (C + 2A) = x^2.$$

Comparing the coefficients of x^2, x^1 and x^0 on both sides, leads to the system of equations $A = 1, B = 0, C + 2A = 0$, which leads to the solution $A = 1, B = 0, C = -2$. Hence $y = x^2 - 2$ is a solution of the given differential equation.

THE METHOD OF UNDETERMINED PARAMETERS IN PROOFS. In many proofs in number theory and elsewhere (e.g. **[I]** pp. 27–28), we take parameters, say t_0 and ρ, fiddle with them, and only at the end commit ourselves to a relation between them (for example, **[I]**, p. 28, $\rho = 1/(\log t_0 + 2)$), that produces the desired effect.

HOW TO GENERALIZE DINITZ'S CONJECTURE?. There is something too narrow and 'square' about the statement of the Dinitz conjecture. A natural generalization that comes to mind is to arbitrary graphs. Calling the elements of the sets $A_{i,j}$ 'colors', the task of the Dinitz conjecture is to color each cell (i, j) by one of the colors of the set of colors $A_{i,j}$ that the cell is allowed to use, in such a way that no two cells sharing the same row, or the same column, can be colored by the same color. This immediately brings to mind graph coloring. The $n \times n$ discrete square is an undirected graph having the n^2 vertices $\{(i, j) | 1 \le i, j \le n\}$, and each vertex (i, j) is connected to the $2(n - 1)$ vertices (i, j'), $j' \neq j$ and (i', j), $i' \neq i$. A natural generalization would have the form

> *If G is any graph in a class X (that includes squares) then whenever each vertex v is assigned a set of colors A_v, and the cardinalities of the sets A_v satisfy condition $Y(G)$, then it is possible to properly color the vertices of G so that the color of each vertex v is drawn from the set A_v.*

(A coloring of a graph is *proper* if two vertices joined by an edge always receive different colors.)

For the time being, both the class X, and the condition Y, are left *blank*. All we need is that the class X contains the graphs of squares and the condition $Y(G)$ becomes 'having cardinalities $\ge n$' when G happens to be the $n \times n$ square.

OCCAM'S RAZOR AND SPECIALIZATION. Properly coloring a graph means that for every two vertices x and y that are connected by an edge, we require that:

The colors assigned to x and y must differ.

This statement really embodies two statements:

The color of x is different from the color of y AND The color of y is different from the color of x.

What a waste! Following Occam's advice, we can drop either one of the two statements. This leads to the idea of directing the edges of our $n \times n$ square-graph and to consider the set of *directed graphs*. This class might be easier to handle, since it has more structure. For any given graph of e edges, there are 2^e ways to make it a directed graph. A proper coloring of a directed graph is assigning colors to each vertex such that whenever there is an edge from vertex x to vertex y, the color assigned to x must differ from the color assigned to y. So in order to properly color an undirected graph, all you need is to be able to color a single one of its many possible directed versions. So now we have one more free parameter at our disposal: the way to 'orient' the graph of the $n \times n$ square. Let's call this orientation Z. The proposed generalization/specialization is now:

> *If G is any directed graph in a class X (that includes the squares with orientation Z) then whenever each vertex v is assigned a set of colors A_v, and the cardinalities of the sets A_v satisfy condition $Y(G)$, then it is possible to properly color the vertices of G so that the color of each vertex v is drawn from the set A_v.*

We would be done if we could find *some* class X, *some* orientation Z, and *some* condition Y, that would enable a proof, such that the squares with orientation Z belong to X, and:

$$Y(n \times n \text{ square with orientation } Z) = \left[|A_{i,j}| \geq n, \text{for all } 1 \leq i, j \leq n\right].$$

Our best bet would be an inductive proof, since graphs are so amenable to induction. Such a proof would presumably consist of a recursive algorithm to color the vertices that would involve, at each step, getting rid of some of the vertices and edges, as well as of some of the colors, thus shrinking the graph, that must stay in our class X, whatever it is, and shrinking the sets A_v, in such a way that condition Y, whatever it is, still holds.

But first let's impose some natural restrictions on the orientation Z of the graph of the square. One of the great principles of mathematics (and life) is *symmetry* and *balance* (e.g., balancing the budget). The number of neighbors of each of the n^2 vertices of the $n \times n$ square is $2n - 2$. When we stick arrows in the edges, it makes sense to do so in such a way that at each vertex there would be as many outgoing edges as incoming edges. So let's impose, tentatively of course, the following condition on the still elusive orientation Z:

> *The orientation Z of the $n \times n$ square should be such that in the resulting directed graph, every vertex has outdegree $n - 1$.*

(The *outdegree* of a vertex is the number of edges coming out of it.)

Now it is time to think of condition Y. The larger the cardinality of the set A_v, the more options we have to color the vertex v. On the other hand, the larger the outdegree of v, the more restrictions we have. Since more freedom should go hand in hand with more responsibility, it makes sense that the condition Y on the cardinality of the set A_v should be related to the outdegree of the vertex v. Since the color of any vertex v should be different from the colors of all its (outgoing) neighbors, that might happen to be all distinct, the number of 'optional colors' at v, i.e. the cardinality of A_v, should be at least one more than the outdegree of v. But wait a minute! In our 'symmetric orientation' Z, the outdegrees are all $n - 1$ and in the statements of Dinitz's conjecture all the cardinalities of the sets A_v are

$\geq n$, one more than the outdegree. This leads us to conjecture that the condition $Y = Y(G)$ should be: $|A_v| \geq outdegree(v) + 1$.

Plugging this (tentative!) condition Y into the 'undetermined generalization' of the Dinitz conjecture, we are lead to the following statement:

If G is any directed graph in a class X (that includes the squares with orientation Z), then whenever each vertex v is assigned a set of colors A_v of cardinality > outdegree(v), it is always possible to properly color the graph in such a way that the color of v is drawn from A_v.

It now remains to find the class X that will make the proof work, and then make sure that there is an orientation Z of the $n \times n$ square such that the outdegree of every vertex is $n - 1$, and that belongs to X.

It is easy to see that the class of *all* directed graphs is too big (why?). On the other extreme the empty class X obviously (and vacuously) satisfies the theorem, but no orientation of the square can ever belong to it, of course.

Anyway, let's leave the nature of the class X blank for now, and try and prove the 'generalized' Dinitz statement. Pick one of the colors in the union of the A_v's; let's call it 'red'. We would like to color 'red' at least one of the vertices that are allowed to be colored 'red', remove these vertices and their incident edges, and thereby get a smaller graph to which we would like to apply induction. In order for the induction to work, the smaller graph G' must still belong to the class X and satisfy condition Y (that the corresponding sets A'_v have cardinality strictly larger then the outdegree of v for every vertex v in the reduced graph G').

When we pick a subset of the vertices to be colored 'red', this subset should be *independent*, i.e., no pair of its members can be connected by an edge; otherwise, the coloring would not be proper.

The vertices that were colored 'red', were among all those vertices v that had the 'red' option, i.e., those for which 'red' $\in A_v$. All the other vertices that had 'red' as one of their options, but were *not* colored 'red', now lose that option. For induction to work, we need that the reduced graph should still satisfy condition Y, which means that these vertices, which are still waiting their turn to be colored, but just lost one of their options, should also lose one of their (outgoing) neighbors. The only way that this could happen is for the 'frustrated red' vertices to have had at least one neighbor amongst the 'departing reds'. Then having colored the 'realized red' vertices 'red', and having removed them, leaves us a graph in which each of the 'frustrated red' vertices gets compensated for its loss of the 'red' option, by getting rid of (at least) one of its annoying (outbound) neighbors.

In order for the difference between the cardinality of the sets A_v and the outdegree of v to be still ≥ 1, we need that out of all the vertices that have 'red' as one of their options, it is possible to pick an *independent* subset of vertices that would exercise that option, in such a way that all the other vertices that had 'red' as one of their options before, but did not use this option, would have an edge leading to one of those vertices that did get colored 'red.' If this is the case, removing the vertices that were just colored 'red', and the edges adjacent to them, would then yield a smaller graph G' that should still belong to X, with correspondingly smaller sets A_v that still satisfy condition Y.

Since we don't know beforehand which of the vertices would have 'red' (or, later, 'green' or any other color) as one of their options, and also want property X to be 'hereditary' (with respect to induced subgraphs), we should 'leave our options open' and require that *any* subset S of vertices should have this property of there

always being an *independent* subset $S' \subset S$ such that every vertex in $S - S'$ has an edge directed toward some vertex of S'. This is exactly the property X that we have been looking for, and the proof that we already have, works with it. So now we can formulate:

Definition. *A directed graph G has property X if for every subset of vertices S there is an independent subset $S' \subset S$ such that every vertex in $S - S'$ has an edge directed toward a vertex of S'.*

We have just proved:

The 'Trivializing' Generalization. *Let G be a directed graph having property X (defined above). If every vertex v is given a set of colors A_v whose cardinality exceeds the outdegree of v, then it is always possible to properly color G in such a way that the color of v is picked from A_v.*

But is this indeed a generalization of the statement of the Dinitz conjecture? We still need to find an orientation Z of the graph of the $n \times n$ square such that every vertex has outdegree $n - 1$, and that has property X.

One of the many possible ways of picking Z is by picking the following orientation. Looking at the Latin square (L) given above, the horizontal (vertical) edges are directed from the smaller (larger) entries to larger (smaller) ones. In other words:

$$(i,j) \to (i',j) \text{ if } [(i+j-2) \bmod n] > [(i'+j-2) \bmod n],$$

$$(i,j) \to (i,j') \text{ if } [(i+j-2) \bmod n] < [(i+j'-2) \bmod n].$$

To prove property X, Galvin invokes the famous Gale-Shapley 'Stable Marriage' theorem (**[GS]**, **[PTW]**), with the rows representing men, the columns representing women, and an arrow from (i,j) to (i',j) meaning that Ms. j prefers Mr. i' to Mr. i while an arrow from (i,j) to (i,j') meaning that Mr. i prefers Ms. j' to Ms. j. Having property X is easily seen to be equivalent to the existence of a stable marriage, even if some of the relationships are removed, because of the laws of the land.[4] In this more general situation, it is no longer guaranteed that everybody gets married, but those who do, do so without fear of being scorned.

I believe that even if the Gale-Shapley algorithm and/or theorem did not exist, it would not have been too hard either to discover it from scratch, or to prove by other means (e.g., induction) that there is some orientation Z (in particular, the one given above), that satisfies property X. We invite the reader to do this right now! □

Postscript. The true story is even more amazing, and I hope that Galvin will write up the story that he told me after he received the first draft of this paper. Since this is *his* story, not mine, I will not give it away, except to quote Noga Alon who said: 'The moral of the (true) story of how Galvin found his proof is not to follow Hilbert, but to follow a simpler adage: *Know where to look things up*'. Of course, just as with Hilbert's advice, this is easier said than done, and it takes someone like Galvin to use this so effectively. It is interesting to note that the right Y and Z were already present in **[AT]** and **[J]**.

[4]For example forbidding (i,j) where Mr. i is a Cohen and Ms. j is a divorcée.

Noga Alon has informed me that Kalai's two-and-a-half page exposé was based on a one-page description that Alon sent Kalai, and that Alon wrote up based on Galvin's letter to him. I wish to thank Mireille Bousquet-Mélou, Fred Galvin, Bruno Salvy, and Herb Wilf for helpful remarks on an earlier version.

Note. John Noonan, of Temple University, has written Maple programs that implement the algorithm in Galvin's proof and the Gale-Shapley algorithm. They are available by anonymous `ftp` to `ftp.math.temple.edu` in directory `pub/noonan`, or via Mosaic to `http://www.math.temple.edu/~noonan`.

REFERENCES

[AT] N. Alon and M. Tarsi, *Coloring and orientations in graphs*, Combinatorica **12** (1992), 125–134.

[ERT] P. Erdös, A. L. Rubin, and H. Taylor, *Choosability in Graphs*, Congr. Numer. **26** (1980), 122–157.

[GS] D. Gale and L. S. Shapley, *College admissions and the stability of marriage*, Amer. Math. Monthly **69** (1962), 9–15.

[G] Fred Galvin, *The list chromatic index of a bipartite multigraph*, J. Combin. Theory Ser. B. **63** (1995), 153–158.

[H] David Hilbert, *Mathematical Problems*, Lecture delivered before the International Congress of Mathematicians at Paris in 1900, Bull. Amer. Math. Soc. **8** (1902), 437–479. [Translation by Mary Winston Newson of the German original that appeared in *Göttinger Nachrichten*, 1900, pp. 253–297.] [Reprinted in: 'Mathematical developments arising from Hilbert Problems', F. E. Browder, Ed., 'Proceedings of Symposia in Pure Mathematics' v.XXVIII-Part 1, Amer. Math. Soc., Providence, 1976, pp. 1–34.]

[I] A. E. Ingham, '*The distribution of prime numbers*', Cambridge University Press, Cambridge, 1932.

[J] Jeannette C. M. Janssen, *The Dinitz problem solved for rectangles*, Bull. Amer. Math. Soc. (N.S.) **29** (1993), 243–249.

[K] Gil Kalai, *E-mail message to Jeannette Janssen* (and possibly to others, but not to me), dated 28 Jan. 1994 11:26 GMT.

[PTW] George Polya, Robert E. Tarjan, and Donald R. Woods, "*Notes on introductory combinatorics*", Birkhauser, Basel and Cambridge, 1983.

Department of Mathematics
Temple University
Philadelphia, PA 19122
zeilberg@math.temple.edu
WWW:http://www.math.temple.edu/~zeilberg
Anon.ftp: ftp.math.temple.edu, directory /pub/zeilberg.

A Note on "Impossible" Paper Folding

Thomas Hull

In [1] the authors draw the line between geometric constructions made with a compass and straight edge and those made by folding a sheet of paper. They wrap things up nicely with the following:

Corollary. *Every thing that is constructible with origami is constructible with a compass and straight edge, but the converse is not true.*

Amazingly enough, and without contradicting the above corollary, we present a paper folding method of trisecting an angle which was developed by Hisashi Abe in the 1970's ([5]).

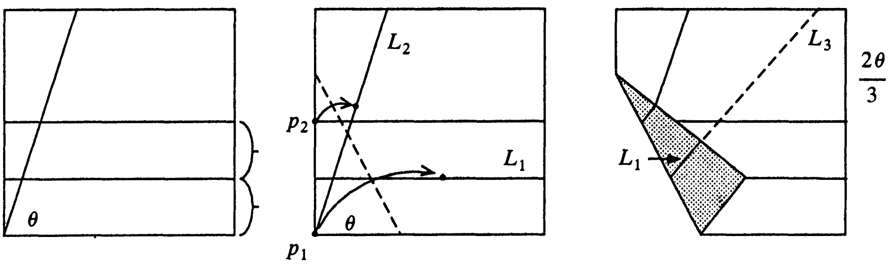


Figure 1.

The proof of this method is a simple exercise in similar triangles. Note that although we assume θ to be acute, this method can be easily extended to obtuse angles.

The reason that this method does not contradict the above Corollary is because in [1] the authors define *origami* (i.e., paper folding) by five "origami axioms." These axioms were selected for their basic simplicity, but don't encompass all possible folding operations. Indeed, step 2 of the above trisection method uses a fold not covered in their axioms. To include this method in [1]'s modeling of paper folding, we would have to include another axiom:

vi) Given distinct points p_1 and p_2 and distinct nonparallel lines L_1 and L_2, there exists a fold that places p_1 onto L_1 and p_2 onto L_2.

In trying to make an axiomatic model of geometric constructions via paper folding, one might want to first look at what origamists have been able to do. In the origami literature there are methods for angle trisection ([2], [7], [8]), doubling cubes ([8], but also see [10]), and folding regular heptagons ([9], [11]), all done using simple folding operations. However, determining the limits of origami constructions is an open problem.

REFERENCES

1. D. Auckly, J. Cleveland, Totally Real Origami and Impossible Paperfolding. *Am. Math. Monthly*, **102** (1995), 215–226.
2. D. Brill, Justin's Origami Trisection, *British Origami*, **107** (1984), 14–15.
3. R. Geretschläger, Euclidean Constructions and the Geometry of Origami (expanded version), unpublished manuscript.
4. E. Frigerio, H. Huzita, A Possible Example of System Expansion in Origami Geometry, in *Proceedings of the First International Meeting of Origami Science and Technology*, ed. H. Huzita (1989), 53–69.
5. K. Husimi, Trisection of angle by H. Abe (in Japanese), in Science of Origami, a supplement to *Saiensu* (the Japanese version of *Scientific American*), Oct. (1980), p. 8.
6. H. Huzita, Axiomatic Development of Origami Geometry, in *Proceedings of the First International Meeting of Origami Science and Technology*, ed. H. Huzita (1989), 143–158.
7. H. Huzita, The Trisection of a Given Angle Solved by the Geometry of Origami, in *Proceedings of the First International Meeting of Origami Science and Technology*, ed. H. Huzita (1989), 195–214.
8. K. Kasahara, *Origami, La Era Neuva* (in Japanese), Sanrio Co., Japan (1989), 262–263.
9. M. Kawamura, *Polyhedral Origami* (in Japanese), μ-Factory (1994), 60–61.
10. P. Messner, Problem 1054 in *Crux Mathematicorum*, **12**, No. 10 (1986).
11. B. Scimemi, Draw of a Regular Heptagon by the Folding, in *Proceedings of the First International Meeting of Origami Science and Technology*, ed. H. Huzita (1989), 71–76.

Department of Mathematics
University of Rhode Island
Kingston, RI 02881-0816
hull@math.uri.edu

"No modern schoolboy can appreciate the blessings which he enjoys in the way of notation till he has seen something of the difficulties with which his predecessors had to wrestle."

M. Barwell (1913).

"In my opinion, a mathematician, in so far as he is a mathematician, need not preoccupy himself with philosophy—an opinion, moreover, which has been expressed by many philosophers. mmmm"

H. Lebesgue, *Scientific American*, September, 1964, p. 129.

A New Minimization Proof for the Brachistochrone

Gary Lawlor

1. INTRODUCTION. The cycloid is the curve traced out by a point on the circumference of a rolling circle. See Figure 1. This curve has two additional names and a lot of interesting history.

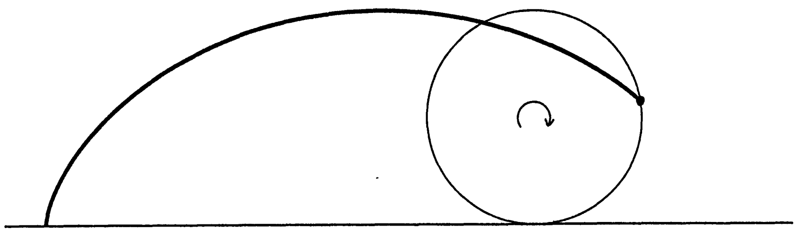


Figure 1. A cycloid, traced out by a fixed point on a rolling circle.

The other two names are the "tautochrone" and the "brachistochrone." In order to talk about the properties of the cycloid that gave it these names, we need to turn the curve upside-down, so that it is concave upward. Thus, throughout this paper, a rolling circle will always be rolling "on the ceiling."

The word "tautochrone" means "same time." In the 1600's Christian Huygens discovered a pendulum whose period is independent of how high the bob swings; this is only approximately true for a regular pendulum.

Huygens found that if obstructions in the shape of half cycloids are placed so that the string of a pendulum wraps around them as it swings (see Figure 2), then the bob also traces out a cycloid. Further, the period of motion is independent of whether the bob makes the full swing or swings only part of the way back and forth. This is equivalent to saying that if we build a ramp in the shape of a cycloid and let a marble roll down it, the time it takes to reach the lowest point is independent of where on the cycloid we started the marble. See Section 8.

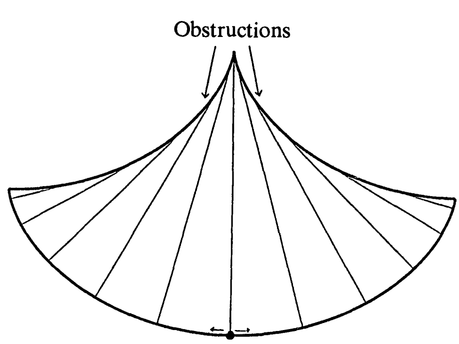


Figure 2. Christian Huygens' pendulum, with period independent of height of swing.

Then in the late 1600's, the mathematician Johann Bernoulli posed a question and invited mathematicians of the time to solve it. He asked, "Let two points A and B be given in a vertical plane. Find the curve that a point M, moving on a path AMB must follow such that, starting from A, it reaches B in the shortest time under its own gravity."

Five prominent mathematicians of the time solved the problem, namely Johann and Jakob Bernoulli, Leibniz, l'Hospital, and Newton. They showed that the solution is also the cycloid, and gave the cycloid its name "brachistochrone," which is Greek for "quickest." The interesting history (including the above quotation from Bernoulli) and an outline of Johann Bernoulli's proof can be found in V. M. Tikhomirov's wonderful book, "Stories about Maxima and Minima" **[T]**.

In this paper we will call the two endpoints P and Q. Bernoulli's "path AMB" we will call a ramp, and the point M we will call a marble.

We will give a new proof that the brachistochrone is the shortest time ramp, using the idea of slicing described in the paper **[L1]**. The philosophy of slicing is to compare two quantities by slicing both into tiny pieces that are easier to compare. We begin with an object that we hope to prove has least volume, area, length, or time among a certain class of competitor objects. We compare this "champion" object with an arbitrary competitor. The goal is to find a strategy for slicing both in such a way that each piece of the champion is smaller or shorter than each piece of the competitor.

The picture of the proof is closely related to Huygens' cycloid pendulum. We will then outline another proof that uses the same basic ideas but demonstrates some additional interesting geometry of the cycloid.

Various modifications of Bernoulli's original question can also be answered by the method of this paper; see Section 9.

We also briefly comment on why the cycloid is the tautochrone.

2. GENERAL FACTS

2.1. Proposition. *The velocity of a marble rolling without friction down a ramp is proportional to* $\sqrt{|y|}$, *if the marble starts at rest at a point where* $y = 0$. *We will choose units so that velocity is* **equal** *to* $\sqrt{|y|}$.

Remark. It is interesting that, in particular, velocity is independent of the shape and length of the ramp and of the x coordinate, and depends only on the y coordinate.

Proof of Proposition 2.1. The kinetic energy of the marble is proportional to mv^2, whether we consider the marble as "sliding" or rolling. The potential energy is mgy. By the law of conservation of energy, the gain in kinetic energy must equal the loss in potential energy. So

$$mg|y| = kmv^2,$$

which means that

$$v = \sqrt{g/k}\sqrt{|y|}. \qquad \blacksquare$$

2.2. Proposition. *The tangent line to a cycloid passes through the lowest point of the rolling circle, and the normal line passes through the highest point.*

Proof: The proof is shown in Figure 3. We draw a vector A tangent to the rolling circle, with the length of A chosen so that it reaches to the horizontal line as in the figure. From its endpoint we draw the horizontal vector B that reaches to the lowest point of the circle. By a general symmetry property of circles, the vectors A and B have the same length. On the other hand, the velocity vector of the point on the rolling circle's circumference can be written as a sum of two vectors, one tangent to the circle (from the rotation) and one horizontal (from the motion of the circle's center). Since the circle rolls without slipping, these two vectors must be of the same length. This means that the velocity vector of the moving point is a multiple of $A + B$, so that the tangent line does pass through the lowest point of the circle.

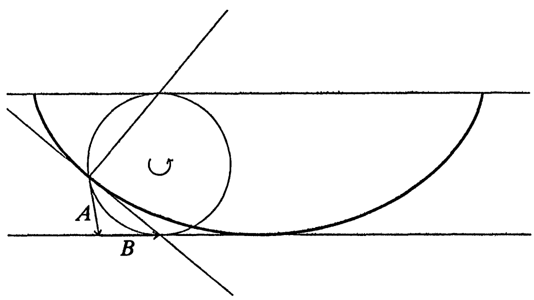


Figure 3. The tangent and normal lines to the cycloid pass through the lowest and highest points of the rolling circle.

Now by a general property of circles, the normal line automatically goes through the opposite point, namely the highest point of the circle. Alternatively, one can rescale A and B into unit vectors, and consider $A^{\perp} + B^{\perp}$. ∎

2.3. Proposition. *In Figure* 4, *the segment KL is perpendicular to the lower cycloid and tangent to the upper cycloid. The point J is the bisector of the segment KL.*

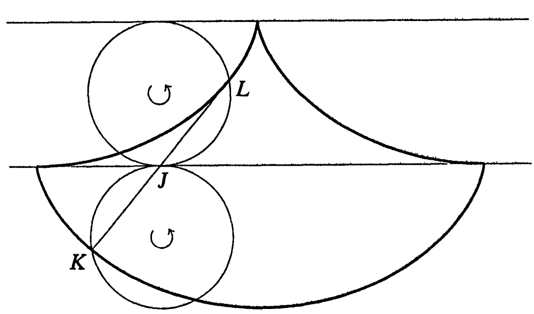


Figure 4. The tangency point J is the midpoint of segment $\overline{KL}$.

Proof: The two circles in the figure are to be thought of as rolling simultaneously to the right, both rolling "on their ceilings" (counterclockwise) and thus slipping against each other rather than rolling like gears. They trace out the upper and lower cycloids. It is interesting to note the similarity between Huygens' tautochrone pendulum and our proof of Bernoulli's brachistochrone property.

In the figure, K is defined as the fixed point on the circumference of the lower rolling circle, and L is the fixed point of the upper circle. Since the circles roll at the same speed, K and L are always opposite each other, with the tangency point J being the midpoint of the line segment between them.

By two applications of Proposition 2.2, since the line segment KL goes through J, the segment is perpendicular to the lower cycloid and tangent to the upper cycloid. ■

3. THE BRACHISTOCHRONE IS THE SHORTEST TIME RAMP. We now state and prove the main theorem.

3.1. Theorem. *Let P and Q be two given points in a vertical plane, with the y coordinate of Q no higher than that of P. Let M be the cycloid whose tangent vector points straight down at point P, and whose generating circle has the right radius so that the cycloid passes through the point Q. For any path from P to Q, consider the time it takes for a marble, starting at rest at P, to roll without friction down to Q. Among all such paths, M is the one on which the marble takes the least time.*

Proof: Our goal in this paper is to present the proof in as elementary a manner as possible. Thus, although each of the following lemmas is true exactly, in proving the second one we will ignore tiny approximation errors. In Section 7 we sketch a rigorous proof by calculus.

Figure 5 is the picture of the proof. We draw narrowly-spaced straight lines perpendicular to the cycloid, and call the lines "strings," reminiscent of Huygens' pendulum. We let P be the highest point of the cycloid on the left, and let Q be any other point of the cycloid. Now every ramp on which a marble may roll from P to Q must cross all of the strings. By Lemma 3.3, the ramp on which a marble travels most quickly from any one string to the next is the cycloid. Thus, by adding up the local results we obtain the theorem. ■

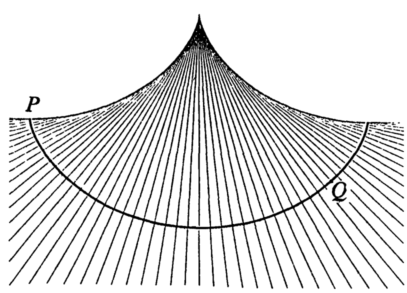


Figure 5. The main idea of the proof that the cycloid is time-minimizing.

3.2. Lemma. *Given a constant $r > 0$, the minimum value of $(r + x)/\sqrt{x}$ occurs where $x = r$.*

Proof: Of course, this is straightforward calculus. Or for a student who has not had calculus, one can write

$$\frac{r+x}{\sqrt{x}} = \frac{r}{\sqrt{x}} + \sqrt{x}.$$

In the latter form, we have a sum of two terms whose product is the constant r. Thinking of the terms as lengths of sides of a rectangle, we see that the area of the rectangle is constant and we want to minimize the (semi)-perimeter. This is accomplished by a square, so we set

$$\frac{r}{\sqrt{x}} = \sqrt{x},$$

so that $x = r$. ∎

3.3. Lemma. *Given any two consecutive strings in Figure 5, for any ramp starting at P consider the tiny increment of time it takes for a marble to cross the gap from the one string to the next. This time increment is always minimized by the cycloid.*

Proof: Pick two consecutive strings, as in Figure 6. There are two factors working against each other in the competition for a fastest path across the gap. Up higher, the gap is narrower, but the speed will be smaller. Very near the top the speed approaches zero but the width does not, so the time will in fact be greater there. Down lower, the speed is greater but the gap is wider. Asymptotically the width grows like y and the speed grows like $\sqrt{y}$, so the time is larger there as well. Thus, somewhere in the middle the crossing time is minimized. Let U be the intersection point of the two chosen strings. Let r be the distance from U to the point V where the first string crosses $y = 0$. Let x be the distance from V to a place where some ramp crosses the gap. Now the velocity is proportional to $\sqrt{x}$, and the width of the gap is proportional to $r + x$, so the time to cross orthogonally is proportional to $(r + x)/\sqrt{x}$, which is minimized when $r = x$, which by Proposition 2.3 is where the brachistochrone crosses the gap. (Note that we have used the approximating assumptions that the speed of the marble is constant as it crosses the gap, that the quickest ramp across the gap is a line segment perpendicular to the first string, and that the two strings intersect at the point where the first one is tangent to the upper cycloid in Figure 4.)

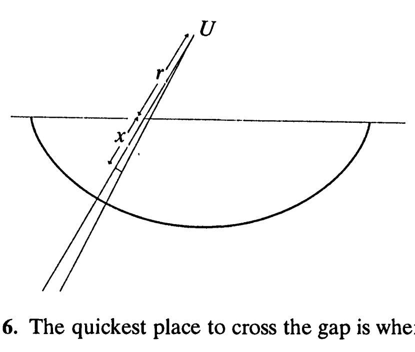


Figure 6. The quickest place to cross the gap is where $x = r$.

4. FINDING THE CYCLOID IN THE FIRST PLACE. We have shown how to prove that a cycloid is the shortest-time ramp. We have not mentioned how to discover the cycloid as a candidate in the first place. This was done by the 17th

century mathematicians by setting up and solving a differential equation. Michael Kerckhove at the University of Richmond had the idea that slicing could also be used as a tool for finding such a differential equation. This approach is proving successful and is a topic of current research.

5. COMPARISON WITH OTHER PROOFS. Recall that the strategy of slicing is to compare two quantities by slicing them into tiny pieces that are easier to compare and for which the desired inequality still holds.

The solutions found in the 17th century by Leibniz and by Johann Bernoulli can both be described as slicing with horizontal lines; see **[T]**, pp. 58–62. The modern technique called the calculus of variations can be viewed as slicing with vertical lines; see L. C. Young's wonderful book **[Y]**. With horizontal and with vertical slicing, it is not true that each piece of the curve will minimize time across its respective gap. Competitor curves will take a shorter time to cross some gaps and a longer time to cross others. In the calculus of variations proof, the extra bookkeeping thus required is accomplished by integration by parts. Leibniz allows only a tiny piece of the curve to vary at a time, and Bernoulli argues by analogy with the refraction of light.

Young is careful to point out that in all of these methods mentioned above a lot of work remains to complete the proof. A complete argument goes as follows: One shows that there does exist a time minimizing curve, and then shows that such a curve (if it is smooth) must satisfy a certain differential equation, and finally finds **all** curves satisfying the differential equation and compares them to see which one takes least time.

In contrast, the method of slicing does not depend on a proof of existence of a minimizing curve. And, although slicing partially localizes the problem, we still keep a sufficiently global view to keep track of all competitors throughout the process.

6. SKETCH OF ANOTHER SLICING PROOF. We now sketch a second proof, which is done by the same basic idea as the first, namely, slice up the plane with curves perpendicular to the brachistochrone ramp, such that the fastest way across the gap between consecutive curves is that taken by the brachistochrone. The geometry is interesting. The result obtained is not quite as strong as before.

For this proof, we can only take half of one cycle of a cycloid, not extending beyond the point where the cycloid becomes horizontal. This time, the slicing curves are themselves cycloids. They have the same size and shape as the brachistochrone cycloid, but are turned right side up (see Figure 7). It is a wonderful fact that no matter how far we shift the cycloids left or right, if they intersect the brachistochrone they do so orthogonally. (The two horizontal lines in Figure 7 are parallel; the apparent narrowing to the right is an illusion.)

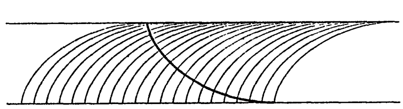


Figure 7. The picture of another slicing proof.

Now consider a fixed gap between consecutive slicing curves. It can be calculated that the shortest time path across the gap is not unique; it takes the same amount of time to cross the gap orthogonally no matter where you cross it.

If S is some other ramp with the same starting and ending points as the (half) brachistochrone, and if S does not go "below ground", i.e., below the level of the endpoint, then S must cross all of the slicing cycloids. The best it can do is to cross all of them orthogonally, which is what the brachistochrone ramp does.

7. PROOF SKETCH OF THEOREM 3.1 USING CALCULUS. One way to include calculus is to define a function $f(x, y)$ by declaring it to equal zero at the top of the brachistochrone ramp, and at each point of the ramp let it equal the time it takes for the marble to get there. Extend the function by letting its level sets be the slicing curves (the straight strings of the first proof or the upside-down cycloids of the second proof). Now prove the inequality $\|\nabla f\| \leq 1/\sqrt{|y|}$ with equality on the brachistochrone. This can be done by implicit differentiation. Let M be the brachistochrone and S a comparison ramp with the same endpoints. Then write the inequalities

$$\begin{aligned}\text{Time}(M) &= f(Q) - f(P) = \int_M \nabla f \cdot \mathbf{ds} = \int_S \nabla f \cdot \mathbf{ds} \\ &\leq \int_S \|\nabla f\|\, ds \leq \int_S \frac{1}{\sqrt{|y|}} ds = \text{Time}(S).\end{aligned}$$

This proof is an example of the method called "calibrations," whose merits were demonstrated in a 1980 landmark paper by Reese Harvey and Blaine Lawson [**HL**]. The method of calibrations (like slicing) provides a global minimization result. Further comments on the comparison between calibrations and slicing are found in the slicing paper [**L1**]. A good exposition on calibrations is the paper [**M1**].

8. THE TAUTOCHRONE. Why does it take the same amount of time for a marble to roll down to the center point C of the ramp, regardless of where we start it? The answer is that the motion is harmonic. If we set a marble at the center point C, there will be no force to move it from this equilibrium point. Now it turns out that if we set a marble at any other point of the ramp and measure its distance from C (measuring arc length along the ramp), then the force acting on the marble (the component of gravity parallel to the ramp) will be proportional to the distance away from C. So if we start marble M some distance away from C, and we simultaneously start marble N (say) twice as far out from C, then marble N will always have twice the force acting on it, so twice the acceleration, so twice the velocity, so it will go twice as far as marble M in the same amount of time, and both will reach C at the same time. The same is true with the word "twice" replaced by "α times as much," for any positive real number α.

9. OTHER QUESTIONS. The geometric proof of the time-minimizing property of the brachistochrone (the one with straight lines) opens the door to the investigation of interesting related problems. Among them are:

(1) What is the quickest path from P to Q if the initial speed at P is nonzero?
(2) What is the quickest path if the cycloid solution is ruled out because it would dip below the level of the floor?
(3) What is the quickest path from a point P to anywhere on a line L?

The first can be answered by calculating how high we should draw the horizontal line $y = 0$. That is, we pretend that a marble has gained its initial velocity by rolling down a ramp from a starting point higher than P. Having drawn the line $y = 0$, we then find the right-sized cycloid that starts vertically at $y = 0$ and then passes through the points P and Q. The proof that this is the best ramp is the same. Draw lines perpendicular to this cycloid; the cycloid is the quickest path across any gap.

The quickest ramp for the second question consists of two pieces of cycloids with a line segment in between. Find the cycloid that is vertical at P and is tangent to the floor. Divide it in half. Move the right-hand piece to the right until it passes through Q. Join with a line segment along the floor. The proof is again by drawing lines perpendicular to the ramp everywhere. The lines perpendicular to the floor segment are parallel; the quickest path across those gaps is along the floor (since you can't go below the floor).

The third question is solved by a cycloid vertical at P and perpendicular to L, with the same proof.

The reader may enjoy devising and solving questions with more complicated obstructions (like question 2) and/or more complicated free boundary (like question 3).

10. FURTHER READING. For further study of current work in geometric measure theory in proving minimization of area, length, time, etc., I recommend the papers **[Bk]**, **[K]**, **[L1]**, **[L2]**, **[LM]**, and **[M2]**.

REFERENCES

[Bk] Ken Brakke, Soap films and covering spaces, Research report, The Geometry Center, University of Minnesota, July, 1993.

[HL] Reese Harvey and H. Blaine Lawson, Jr., Calibrated Geometries, Acta Mathematica **148** (1982), 47–157.

[K] Michael Kerckhovė, Isolated orbits of the adjoint action and area-minimizing cones, Proceedings of the AMS **121** (1994) 497–503.

[L1] Gary Lawlor, Proving area-minimization by slicing, preprint.

[L2] Gary Lawlor, A sufficient criterion for a cone to be area-minimizing, Memoirs Amer. Math. Soc. **91**, No. 446 (1991).

[LM] Gary Lawlor and Frank Morgan, Paired calibrations applied to soap films, immiscible fluids, and surfaces or networks minimizing other norms, Pacific Journal of Mathematics, **166 no. 1** (1994), 55–83.

[M1] Frank Morgan, Area-minimizing surfaces, faces of Grassmannians, and calibrations, American Math. Monthly **95 no. 9** (1988), 813–822.

[M2] Frank Morgan, Soap films and mathematics, Proc. Symp. Pure Math. **54** (1993), 375–380.

[T] V. M. Tikhomirov, *Stories about Maxima and Minima. Translated from the Russian by Abe Shenitzer.*, Amer. Math. Soc. and Math. Assoc. of America, 1990.

[Y] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders Co., Philadelphia, 1969.

Department of Mathematics
Brigham Young University
Provo, UT 84602
lawlor@math.byu.edu

NOTES

Edited by: **John Duncan**

Strategies for the Shannon Switching Game

Richard Mansfield

We present a proof that the Shannon switching game on a graph with distinguished vertices A and B has a winning strategy for player Short iff the graph has a subgraph connecting A to B with two edge disjoint spanning trees. This theorem was first proved by Lehman in **[11]**. The difficulty is that his proof is phrased in terms of matroids and appears very difficult. However, I was able to cull a neat and short but elegant proof out of Lehman's work and that is what I want to present here. Using this proof, we will be able to give a strategy for player Cut for the graphs in which he has a win. This strategy is not quite as neat as Short's strategy, and does not appear to be computationally feasible, but at least it does not involve any look ahead analysis.

We are given a multi-graph G with two distinguished vertices, A and B. The two players, Cut and Short play alternately. Short's move consists of marking an edge while Cut may remove any unmarked edge from the graph. If Short can succeed at marking an entire path from A to B, he wins. Otherwise Cut wins. As we shall see shortly, if G contains two trees connecting A to B having no edges in common but sharing the same vertex set, then Short has a winning strategy even if he goes second. Our main theorem is the converse, that if he has a winning strategy from the second position, then there must be two such trees. Brualdi has written an excellent expository article about the switching game. See **[1]**. This article has been essentially repeated as a section in his book [**2**, Section 11.6].

Let us say that a graph is *positive* if it has an edge disjoint pair of spanning trees. Then what we wish to show is that G has a positive subgraph containing both A and B if and only if Short has a winning strategy even if he plays second. The reader may easily verify that the union of two positive graphs is either disconnected or positive.

The reader may also easily verify that if there is a positive subgraph connecting A to B, then Short has a winning strategy. He proceeds as follows: If Cut removes an edge from one of the two trees, Short finds an edge in the other tree which reconnects the broken tree and marks it. He then has two spanning trees for the subgraph with only marked edges in common.

Let us prove the converse. Suppose Short goes second and has a winning strategy. We must prove the existence of a positive subgraph. So Cut goes first and deletes an edge. Short then marks an edge, call it a. By induction, we may assume that there are two trees, S_1 and T_1 from A to B having a common vertex set but only the edge a in common.

We now need a lemma. We shall state and use the lemma and only when the main proof is finished will we come back and prove it. The lemma is the following:

If S and T are two trees with a common vertex set and at most one edge in common and if P and Q are two distinct vertices of these trees, then either $S \cup T$ has a positive subgraph connecting P to Q or it has two spanning trees with only one edge in common having the additional property that in at least one of the trees the common edge lies on the path from P to Q. As a first use of this lemma, we may as well assume that a lies on the path in S_1 from A to B.

If every edge lying on the path from A to B in T_1 was spanned by a positive subgraph of G, then the union of these positive subgraphs would be a positive graph connecting A to B. So choose an edge b not spanned by any positive subgraph of G and lying on the path in T_1 from A to B. Consider the fact that Cut could have chosen b on his first move instead of the one he did choose. Again, by induction, we get trees S_2 and T_2 avoiding b with a common vertex set and at most one edge in common joining A to B.

Deleting the edge a from S_1 splits S_1 into two connected components with A in one of the components and B in the other. Deleting the edge b from T_1 has a similar effect. Thus, since S_2 and T_2 are both connected graphs containing both A and B, the graphs $T_2 \cup (T_1 \setminus \{b\})$ and $S_2 \cup (S_1 \setminus \{a\})$ are both connected. In order to apply the Lemma, we also need the condition that the graphs have only one edge in common. For these graphs this need not be the case since S_1 may overlap T_2 and T_1 may overlap S_2. However, since T_2 spans every edge in S_2, we may further take all the edges of $S_2 \setminus T_2$ away from $T_2 \cup (T_1 \setminus \{b\})$ and still have a connected graph. Similarly, we may delete $T_2 \setminus S_2$ from $S_2 \cup (S_1 \setminus \{a\})$ without losing connectivity. Thus, the two graphs, $T_3 = T_2 \cup (T_1 \setminus (S_2 \cup \{b\}))$ and $S_3 = S_2 \cup (S_1 \setminus (T_2 \cup \{a\}))$ are both connected. Both these graphs span the edge b, but neither contains b. Furthermore, they have at most one edge in common. Using the lemma again, we see that $S_3 \cup T_3$ must have two spanning trees whose only common edge lies on the path in one of the two trees between the two ends of b. That tree may then be disconnected by deleting the common edge from it and then reconnected by adding the edge b. This proves the theorem and we are left with proving the lemma.

To this end, let S and T be the two given trees. A *spanning chain* is a sequence $C_0, C_1, \dots, C_n$ such that C_0 is the path in S from P to Q and for each $i < n/2$, the path C_{2i+1} is the path in T spanning an edge in C_{2i} and C_{2i+2} is the path in S spanning an edge in C_{2i+1}. If the common edge is not on any path in any spanning chain, then let S' be all the edges e in S for which there is a spanning chain with a path containing e, and let T' be all the edges e in T for which there is a spanning chain with a path containing e. We can easily see that S' and T' are two edge disjoint trees with a common vertex set connecting the two given vertices and thus the first possibility holds.

We complete the lemma by induction on the length of the shortest chain containing the common edge. Let $C_0, \dots, C_{n-1}, C_n$ be such a shortest chain. If n is 0, there is nothing to prove. If n is even, then C_n is a path in S, otherwise it is a path in T. In any event, the common edge lies in C_n. Proceed as follows: Delete the common edge from the appropriate tree. (S if n is even, T if n is odd.) Then reconnect the tree by adding to it the edge in C_{n-1} spanned by C_n. This new doubled edge lies on a shorter spanning chain and so the lemma follows by induction.

The above proof yields a strategy for player Cut. Suppose it is Cut's turn and Short does not have a forced win. Cut should then find a pair of trees S and T connecting A to B with a common vertex set but only one unmarked edge in common. If there are no such trees, he may play at random. Finding such a pair

however, he should perform the construction given in the lemma to ensure that the common edge lies on the path in S from A to B. He should then find an edge on the path in T from A to B which is not in any positive subgraph of G and remove it. Our proof has just shown that if this procedure allows Short to get a forced win, then Short already had a forced win before Cut made his move.

REFERENCES

1. Richard A. Brualdi, *Networks and the Shannon switching game*, Delta (Waukesha) 4 (1974), 1–23.
2. ______, *Introductory combinatorics*, North Holland, New York, 1992.
3. John Bruno and Louis Weinberg, *A constructive graph-theoretic solution of the Shannon switching game*, IEEE Trans. Circuit Theory CT-17 (1970), 74–81.
4. Shimon Even and R. Endre Tarjan, *A polynomial problem which is complete in polynomial space*, Proceedings of the Seventh Annual ACM Symposium on the Theory of Computing, vol. 7, 1975, pp. 66–71.
5. Harold N. Gabow, *Forests, frames, and games: algorithms for matroidal sums and applications*, Algorithmica 7 (1992), 465–497.
6. Yahya Ould Hamidoune, *The directed Shannon switching game and the one-way game*, Graph theory with applications to algorithms and computer science (New York), Wiley, New York, 1984, pp. 391–400.
7. ______, *Jeux de commutation orientes sur les graph et les matroids*, C. R. Acad. Sci. Paris Ser. I Math. 298 (1984), 497–499.
8. ______, *Directed switching games on graphs and matroids*, J. Combin. Theory Ser. B 40 (1986), 237–269.
9. ______, *A solution to the box game*, Discrete Math. 65 (1987), 157–171.
10. ______, *A solution to the misere Shannon switching game*, Discrete Math 72 (1988), 163–166.
11. Alfred Lehman, *A solution of the Shannon switching game*, J. Soc. Indust. Appl. Math. 12 (1964), 687–725.
12. Yohei Yamasaki, *Shannon switching games without terminals. II*, Graphs and Combinatorics 5 (1989), 275–282.
13. ______, *Shannon switching games without terminals. III*, Graphs and Combinatorics 8 (1992), 291–297.
14. ______, *Shannon-like games are difficult*, Discrete Math. 111 (1993), 481–483.

Department of Mathematics
Penn State University
University Park, PA 16802
melvin@math.psu.edu

k-volume in $\mathbf{R}^n$ and the Generalized Pythagorean Theorem

Gerald J. Porter

Let $V(\mathbf{u1}, \ldots, \mathbf{uk})$ be the parallelepiped generated by k vectors, $\mathbf{u1}, \ldots, \mathbf{uk}$ in $\mathbf{R}^n$. What is the k-volume* of the parallelepiped? If $k = n$, then every linear algebra student knows that the volume is the determinant of the matrix whose columns (or

*Volume may be defined either axiomatically or inductively. For an axiomatic definition see Chapter 5, Section 19 of Curtis, *Linear Algebra*, Springer-Verlag, 1984. For an inductive definition see Birkhoff and MacLane, *A Survey of Modern Algebra*, Macmillan, 1957.

rows) are the vectors, $\mathbf{u1}, \ldots, \mathbf{un}$. If $k = 1$, volume is length and the length satisfies the Pythagorean Theorem, $|\mathbf{u}|^2 = (\mathbf{u}_1)^2 + (\mathbf{u}_2)^2 + \cdots + (\mathbf{u}_n)^2$, where $\mathbf{u}_i$ is the ith coordinate of the vector $\mathbf{u}$. Our goal in this note is to derive two formulas for the volume that generalize these "extreme" cases.

Generalization of the Determinant Solution. If $k = 2$, volume is area and

$$\mathbf{Volume} = |\mathbf{u1}| \cdot |\mathbf{u2}| \cdot \mathbf{sin}\left(\mathbf{acos}\left(\frac{\mathbf{u1} \cdot \mathbf{u2}}{|\mathbf{u1}| \cdot |\mathbf{u2}|}\right)\right)$$

$$\mathbf{Volume} = |\mathbf{u1}| \cdot |\mathbf{u2}| \cdot \frac{\sqrt{(|\mathbf{u1}|)^2 \cdot (|\mathbf{u2}|)^2 - (\mathbf{u1} \cdot \mathbf{u2})^2}}{(|\mathbf{u1}| \cdot |\mathbf{u2}|)}$$

$$\mathbf{Volume}^2 = (|\mathbf{u1}|)^2 \cdot (|\mathbf{u2}|)^2 - (\mathbf{u1} \cdot \mathbf{u2})^2.$$

This is nothing other than the determinant

$$\left|\begin{pmatrix} \mathbf{u1} \cdot \mathbf{u1} & \mathbf{u1} \cdot \mathbf{u2} \\ \mathbf{u1} \cdot \mathbf{u2} & \mathbf{u2} \cdot \mathbf{u2} \end{pmatrix}\right|$$

of the matrix product

$$\begin{pmatrix} \mathbf{u1}^T \\ \mathbf{u2}^T \end{pmatrix} \cdot \begin{pmatrix} \mathbf{u1} & \mathbf{u2} \end{pmatrix} = \begin{pmatrix} \mathbf{u1} & \mathbf{u2} \end{pmatrix}^T \cdot \begin{pmatrix} \mathbf{u1} & \mathbf{u2} \end{pmatrix}$$

where $(\mathbf{u1} \quad \mathbf{u2})$ denotes the k by 2 matrix whose columns are $\mathbf{u1}$ and $\mathbf{u2}$. This leads to the following generalization:

Theorem 1. *Let* $\mathbf{u1}, \ldots, \mathbf{uk}$ *be* k *vectors in* $\mathbf{R}^n$. *Then the volume of the parallelepiped generated by the vectors is the square root of the determinant of* $\mathbf{U}^T \cdot \mathbf{U}$, *where* $\mathbf{U}$ *is the matrix whose columns are the vectors,* $\mathbf{u1}, \ldots, \mathbf{uk}$.†

Proof: If the vectors $\mathbf{u1}, \ldots, \mathbf{uk}$ are orthogonal then the volume is

$$|\mathbf{u1}| \cdot |\mathbf{u2}| \cdots |\mathbf{uk}|.$$

In this case,

$$(\mathbf{U}^T \cdot \mathbf{U})_{i,j} = \mathbf{ui} \cdot \mathbf{uj}.$$

Since the $\mathbf{u}$'s are assumed to be orthogonal, it follows that

$$|\mathbf{U}^T \cdot \mathbf{U}| = (|\mathbf{u1}|)^2 \cdot (|\mathbf{u2}|)^2 \cdots (|\mathbf{uk}|)^2 = \mathbf{Volume}(V(\mathbf{u1}, \ldots, \mathbf{uk}))^2.$$

If the vectors are not orthogonal, we may transform them into an orthogonal set by using the Gram-Schmidt procedure. This is applied inductively and the resulting set of orthogonal vectors, $\mathbf{w1}, \ldots, \mathbf{wk}$ is defined by

$$\mathbf{wj} = \mathbf{uj} - \sum_{\mathbf{i}=1}^{\mathbf{j}-1} \frac{\mathbf{uj} \cdot \mathbf{wi}}{\mathbf{wi} \cdot \mathbf{wi}} \cdot \mathbf{wi}.$$

This transformation does not change volume since $\mathbf{wj}$ is the component of $\mathbf{uj}$ that is orthogonal to $\mathbf{w1}, \ldots, \mathbf{w(j-1)}$. Since each of $\mathbf{w1}, \ldots, \mathbf{w}(j-1)$ lies in the span of $\mathbf{u1}, \ldots, \mathbf{u}(j-1)$ it follows that the matrix, $\mathbf{A}$, that transforms the $\mathbf{u}$'s to $\mathbf{w}$'s is upper triangular with 1's on the diagonal and therefore has determinant 1. Since the

†We thank Ladnor Geissinger for pointing out that this theorem can be found, with a different proof, as Theorem 7 of Section 10.3 of Birkhoff and MacLane, *A Survey of Modern Algebra*.

columns of $\mathbf{U} \cdot \mathbf{A}$ are orthogonal, the desired volume is the square root of the determinant of $(\mathbf{U} \cdot \mathbf{A})^T \cdot (\mathbf{U} \cdot \mathbf{A})$ and the theorem follows from the following calculation.

$$\begin{aligned}|(\mathbf{U} \cdot \mathbf{A})^T \cdot (\mathbf{U} \cdot \mathbf{A})| &= |\mathbf{A}^T \cdot (\mathbf{U}^T \cdot \mathbf{U}) \cdot \mathbf{A}| = |\mathbf{A}^T| \cdot |\mathbf{U}^T \cdot \mathbf{U}| \cdot |\mathbf{A}| \\ &= (|\mathbf{A}|)^2 \cdot |\mathbf{U}^T \cdot \mathbf{U}| = |\mathbf{U}^T \cdot \mathbf{U}|.\end{aligned}$$

The Generalized Pythagorean Theorem. Let $\mathbf{e_1}, \ldots, \mathbf{e_n}$ be the standard basis for $\mathbf{R}^n$ (i.e., $\mathbf{e_j}$ is the vector whose jth component is 1 and whose ith component is 0 for i not equal to j).

Let $\Gamma(k, n)$ be the set of increasing sequences, $1 \leq i1 < i2 < \cdots < ik \leq n$.

For $\gamma = (i1 \quad i2, \ldots, ik) \in \Gamma(k, n)$, let $\mathbf{e}_\gamma$ be the k-tuple of basis vectors $\mathbf{e}_{i1}, \mathbf{e}_{i2}, \ldots, \mathbf{e}_{ik}$; $\mathbf{R}_\gamma$ the subset of $\mathbf{R}^n$ spanned by $\mathbf{e}_\gamma$; and $\mathbf{P}_\gamma$ the projection from $\mathbf{R}^n$ to $\mathbf{R}_\gamma$.

The Pythagorean Theorem can be written as:

$$(|\mathbf{v}|)^2 = \sum_{(\Gamma(\mathbf{1}, \mathbf{n}))} (|\mathbf{P}_\gamma \cdot \mathbf{v}|)^2.$$

We generalize this as follows:

Theorem 2. *Let* $\mathbf{u1}, \ldots, \mathbf{uk}$ *be vectors in* $\mathbf{R}^n$ *and let* $V = V(\mathbf{u1}, \ldots, \mathbf{uk})$ *be the parallelepiped generated by the vectors* $\mathbf{u1}, \ldots, \mathbf{uk}$. *Then*

$$(\mathbf{Volume(V)})^2 = \sum_{(\Gamma(\mathbf{k}, \mathbf{n}))} (\mathbf{Volume}(\mathbf{P}_\gamma \cdot \mathbf{V}))^2.$$

In words, the volume squared of a k-parallelepiped is the sum of the squares of the volumes of its projections into the distinct k-dimensional subspaces spanned by the canonical basis.

Theorem 2 follows from the following calculations.

Let $\mathbf{A}$ be an n by k matrix and let $\mathbf{A}^{\langle i \rangle}$ be the ith column of $\mathbf{A}$, then

$$|\mathbf{A}^T \cdot \mathbf{A}| = \mathbf{DET}(\mathbf{A}^T \cdot \mathbf{A}^{\langle 1 \rangle}, \mathbf{A}^T \cdot \mathbf{A}^{\langle 2 \rangle}, \ldots, \mathbf{A}^T \cdot \mathbf{A}^{\langle k \rangle}),$$

where the right-hand side represents the determinant of a matrix as a k-linear function of the columns of the matrix. Let $\mathbf{e_j}$ be the jth element of the canonical basis for $\mathbf{R}^n$. Then the right-hand side above equals

$$\begin{aligned}&\mathbf{DET}\left(\mathbf{A}^T \cdot \sum_{j1=1}^{n} \mathbf{a}_{j1,1} \cdot \mathbf{e}_{j1}, \mathbf{A}^T \cdot \sum_{j2=1}^{n} \mathbf{a}_{j2,2} \cdot \mathbf{e}_{j2}, \ldots, \mathbf{A}^T \cdot \sum_{jk=1}^{n} \mathbf{a}_{jk,k} \cdot \mathbf{e}_{jk}\right) \\ &\quad = \sum_{j1=1}^{n} \sum_{j2=1}^{n} \cdots \sum_{jk=1}^{n} \mathbf{a}_{j1,1} \cdot \mathbf{a}_{j2,2} \cdot \ldots \cdot \mathbf{a}_{jk,k} \cdot \mathbf{DET}(\mathbf{A}^T \cdot \mathbf{e}_{j1}, \ldots, \mathbf{A}^T \cdot \mathbf{e}_{jk}).\end{aligned}$$

Let $\mathbf{A}_j$ be the jth row of $\mathbf{A}$. Since $\mathbf{A}^T \cdot \mathbf{e}_j = \mathbf{A}_j^T$, this equals

$$\sum_{j1=1}^{n} \sum_{j2=1}^{n} \cdots \sum_{jk=1}^{n} \mathbf{a}_{j1,1} \cdot \mathbf{a}_{j2,2} \cdot \ldots \cdot \mathbf{a}_{jk,k} \cdot \mathbf{DET}(\mathbf{A}_{j1}^T, \ldots, \mathbf{A}_{jk}^T).$$

If $jr = js$ for any two indices above, the determinant is zero, so it suffices to take the sum over distinct k-tuples $(j1, \ldots, jk)$.

For $\gamma = (i1, \ldots, ik) \in \Gamma(k, n)$, let $\mathbf{A}_\gamma$ be the k by k submatrix of $\mathbf{A}$ whose jth row is $\mathbf{A}_{ij}$.

Let $\sigma(j1,\ldots,jk)$ be the inverse of the permutation that permutes $(j1,\ldots,jk)$ into an increasing k-tuple $(i1,\ldots,ik)$. Since $\operatorname{sgn}(\sigma)=\operatorname{sgn}(\sigma^{-1})$,

$$\mathbf{DET}\left(\mathbf{A}_{j1}^{T},\ldots,\mathbf{A}_{jk}^{T}\right)=\mathbf{sgn}(\sigma(\mathbf{j1},\ldots,\mathbf{jk}))\cdot\mathbf{DET}\left(\mathbf{A}_{i1}^{T},\ldots,\mathbf{A}_{ik}^{T}\right)$$

and we may rewrite the above expression as

$$\sum_{\Gamma(k,n)}\left(\sum\mathbf{sgn}(\sigma(\mathbf{j1},\ldots,\mathbf{jk}))\cdot\mathbf{a}_{j1,1}\cdot\mathbf{a}_{j2,2}\cdots\mathbf{a}_{jk,k}\right)\cdot\mathbf{DET}\left(\mathbf{A}_{i1}^{T},\ldots,\mathbf{A}_{ik}^{T}\right)$$
$$=\sum_{\Gamma(k,n)}\left(|\mathbf{A}_{\gamma}|\right)^{2}.$$

Let $\mathbf{A}$ and $\mathbf{B}$ be n by k matrices. The above calculation shows that

$$|\mathbf{A}^{T}\cdot\mathbf{B}|=\sum_{(\Gamma(k,n))}|\mathbf{A}_{\gamma}|\cdot|\mathbf{B}_{\gamma}|.$$

This result is essentially the same as the following, which appears on page 212 of Muir: *A Treatise on the Theory of Determinants* (1933).

> If there be two sets of elements both consisting of m rows of n elements (n being greater than m) and a determinant be formed from them in the way in which the product of two determinants is formed, multiplying row by row, then this determinant is equal to the sum of every product whose first factor is a determinant obtained by taking m columns of the first set of elements, and whose other factor is the determinant obtained by taking the corresponding m columns of the second set.

The proof of Theorem 2 can also be given by considering the vector space, $\Lambda^{k}(\mathbf{R}^{n})$, of alternating k-linear functionals on $\mathbf{R}^{n}$. (For details on this and on what follows see Section 5.6 of Hoffman & Kunze, *Linear Algebra*, Prentice-Hall 1961.)

Let $\mathbf{e}_{\gamma}$ be as above. A direct calculation shows that each element of $\Lambda^{k}(\mathbf{R}^{n})$ is determined by its values on the k-tuples $\mathbf{e}_{\gamma}$ for $\gamma\in\Gamma$.

Set $\phi\gamma$ equal to the element of $\Lambda^{k}(\mathbf{R}^{n})$ that equals 1 on $\mathbf{e}_{\gamma}$ and 0 on $\mathbf{e}_{\alpha}$ for $\alpha\neq\gamma$. The set of $\phi\gamma$ for $\gamma\in\Gamma$ form a basis for $\Lambda^{k}(\mathbf{R}^{n})$.

$\Lambda^{k}(\mathbf{R}^{n})$ is an inner product space with a product given as follows:

$$\alpha\cdot\beta=\sum_{\gamma}\alpha(\mathbf{e}_{\gamma})\cdot\beta(\mathbf{e}_{\gamma}).$$

With this inner product the set of $\phi\gamma$, for $\gamma\in\Gamma$, form an orthonomal basis for $\Lambda^{k}(\mathbf{R}^{n})$. Let $\mathbf{A}$ be an n by k matrix. Define a k-linear functional, α, on $\mathbf{R}^{n}$ by

$$\alpha(\mathbf{u1},\ldots,\mathbf{uk})=|\mathbf{A}^{T}\cdot\mathbf{u1},\ldots,\mathbf{A}^{T}\cdot\mathbf{uk}|=|\mathbf{A}^{T}\cdot\mathbf{U}|\text{ where }\mathbf{U}^{\langle j\rangle}=\mathbf{uj}.$$

As functionals we have $\alpha=\sum_{\gamma}\alpha(\mathbf{e}_{\gamma})\cdot\phi\gamma=\sum_{\gamma}|\mathbf{A}_{\gamma}|\cdot\phi\gamma$ where, if $\gamma=(i1,i2,\ldots,ik)\in\Gamma(k,n)$, $\mathbf{A}_{\gamma}$ is the k by k submatrix of $\mathbf{A}$ whose rows are $\mathbf{A}_{i1},\ldots,\mathbf{A}_{ik}$. A direct calculation that shows that $\phi\gamma(\mathbf{u1},\mathbf{u2},\ldots,\mathbf{uk})$ equals the determinant of $\mathbf{U}_{\gamma}$. Setting $\mathbf{A}=\mathbf{U}$, Theorem 2 follows.

Example 1.

$$\mathbf{U}:=\begin{bmatrix}2&-6&4\\6&2&0\\0&1&2\\1&-1&-2\\5&3&-1\end{bmatrix}.$$

The square of the volume of the parallelepiped spanned by the columns of $\mathbf{U}$ equals:

$$|\mathbf{U}^T \cdot \mathbf{U}| = 43641.$$

There are (5 choose 3) = 10 3-dimensional subspaces.

We compute the volume of the projections into these using the ordering (1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 3, 4), (1, 3, 5), (1, 4, 5), (2, 3, 4), (2, 3, 5), (2, 4, 5), (3, 4, 5). Let $\mathbf{I}$ be the five by five identity matrix.

$$\mathbf{P}(\mathbf{i},\mathbf{j},\mathbf{k}) := \text{augment}(I^{\langle i \rangle}, \text{augment}(I^{\langle j \rangle}, I^{\langle k \rangle}))^T$$

$$\mathbf{v}_1 := |\mathbf{P}(1,2,3) \cdot \mathbf{U}| \quad \mathbf{v}_2 := |\mathbf{P}(1,2,4) \cdot \mathbf{U}| \quad \mathbf{v}_3 := |\mathbf{P}(1,2,5) \cdot \mathbf{U}|$$

$$\mathbf{v}_4 := |\mathbf{P}(1,3,4) \cdot \mathbf{U}| \quad \mathbf{v}_5 := |\mathbf{P}(1,3,5) \cdot \mathbf{U}| \quad \mathbf{v}_6 := |\mathbf{P}(1,4,5) \cdot \mathbf{U}|$$

$$\mathbf{v}_7 := |\mathbf{P}(2,3,4) \cdot \mathbf{U}| \quad \mathbf{v}_8 := |\mathbf{P}(2,3,5) \cdot \mathbf{U}| \quad \mathbf{v}_9 := |\mathbf{P}(2,4,5) \cdot \mathbf{U}|$$

$$\mathbf{v}_{10} := |\mathbf{P}(3,4,5) \cdot \mathbf{U}|$$

$$\mathbf{v}^T = (104, -112, -8, -16, -94, 100, 4, -22, 24, 7)$$

$$\mathbf{v} \cdot \mathbf{v} = 43{,}641.$$

Example 2. Let $\mathbf{u1}, \mathbf{u2}, \dots, \mathbf{u}(n-1)$ be vectors in $\mathbf{R}^n$. The definition of the cross-product of two vectors in $\mathbf{R}^3$ can be generalized to a product on $n-1$ vectors in $\mathbf{R}^n$. The resulting product is then orthogonal to each of the $(n-1)$ vectors and has length equal to the $(n-1)$-volume of the parallelepiped generated by $\mathbf{u1}, \dots, \mathbf{u}(n-1)$.

Let $\mathbf{u1}, \dots, \mathbf{u}(n-1)$ be the columns of the matrix $\mathbf{U}$ given below.

$$\mathbf{u} := \begin{bmatrix} 2 & 1 & 4 & 0 & 3 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 2 & 3 & 0 & 7 & 2 \\ -8 & 5 & 7 & 6 & 0 & -1 \\ 2 & -3 & 2 & 3 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 3 & 2 & 0 & -2 & 0 & 3 \end{bmatrix}.$$

The cross-product is

$$\mathbf{v} := \begin{bmatrix} 1268 \\ -2377 \\ -28 \\ -6 \\ -669 \\ -1231 \\ -406 \end{bmatrix}.$$

We leave it to the reader to verify that $\mathbf{v}$ is orthogonal to the column space of U and has length equal to the square root of

$$|\mathbf{U}^T \cdot \mathbf{U}| = 9{,}386{,}531.$$

Department of Mathematics
University of Pennsylvania
209 S. 33rd Street
Philadelphia, PA 19104-6394
gjporter@math.upenn.edu

THE EVOLUTION OF...

Edited by: **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

A Few Expository Mini-Essays

Abe Shenitzer

Sometimes it is possible to describe a mathematical issue of great importance very briefly. This is exemplified by the essays below. (Of the six essays in this paper, two are quotes and two more are essentially quotes.)

(A) Polynomial Equations. Interest in the solution of polynomial equations has been a constant feature of mathematics from the time of the Babylonians to this day. In fact, it is safe to say that until the revolution wrought in mathematics by the ideas of Galois, algebra was synonymous with the solution of equations. The first substantial elaboration of Galois' ideas had to wait until 1870, when Jordan published his influential *Treatise*. Since then algebra has largely concerned itself with the study of structures such as groups, fields, rings, vector spaces, etc. Back to polynomial equations.

Why equations? Granted an interest in functions, it is natural to ask of any function f what is its range and what is the preimage $f^{-1}(a)$ of an element a in the codomain of f. The latter question asks for the solution of the equation $f(x) = a$. This can be normalized as $\phi(x) = f(x) - a = 0$. It is obtaining an explicit solution of this problem for a polynomial which has occupied mathematicians for some 4000 years, and which has been (almost) completely solved only in 1984 by the Japanese mathematician Hiroshi Umemura.

Umemura's achievement seems to be a well-kept secret. I found out about it by reading (parts of) a remarkable book-length essay (172 pages) titled *Mathematics and Physics* by Krzysztof Maurin [1]. Here is how Maurin leads up to Umemura's result:

> Some 2000 years BC the Babylonians could solve certain quadratic equations (so too could the Chinese and the Indians). In the 16th century the Italians (del Ferro, Cardano) solved cubic and quartic equations by extraction of roots. All attempts to obtain similar solutions of the general quintic equation failed. (That they were bound to fail follows from the Abel-Ruffini theorem (1826) which established the impossibility of the solution of the general quintic equation by means of extraction of roots.) It was Galois who stated a group-theoretic condition for the solvability of an n-th order equation by means of extraction of roots. (The solvability of an algebraic equation is equivalent to the solvability of its Galois group.) As early as 1858 Hermite and Kronecker showed (independently) that the quintic equation could be solved by using an elliptic *modular function*. Since

$$(*) \qquad \sqrt[n]{a} = \exp\left(\frac{1}{n}\log a\right) = \exp\left(\frac{1}{n}\int_1^a x^{-1}\,dx\right),$$

extraction of roots involves integration of the function $1/x$ and the use of the exponential function. The idea of Hermite and Kronecker was the following: to solve a fifth-degree equation it is necessary to replace the exponential function exp in $(*)$ by another transcendental function (by an elliptic modular function), and the integral by elliptic integrals. Kronecker thought that the roots of an arbitrary algebraic equation could be found in a similar way. Umemura showed that Kronecker's conjecture was correct: Every algebraic equation has a root that can be expressed in terms of a modular function and hyperelliptic integrals.... .

Note. The rest of Maurin's, quite technical, discussion of Umemura's theorem, not included here, contains an expression for a root of the general n-th order equation. Those interested in this material can obtain copies by getting in touch with Abe Shenitzer. (E-mail: ⟨shenitze@mathstat.yorku.ca⟩)

(B) A Capsule History of Dynamics. This is a quote from pp. 2–3 in: Strogatz, NONLINEAR DYNAMICS AND CHAOS, ©1994 by Addison-Wesley Publishing Company, Inc. Reprinted by permission of the Publisher.

Although dynamics is an interdisciplinary subject today, it was originally a branch of physics. The subject began in the mid-1600s, when Newton invented differential equations, discovered his laws of motion and universal gravitation, and combined them to explain Kepler's laws of planetary motion. Specifically, Newton solved the two-body problem—the problem of calculating the motion of the earth around the sun, given the inverse square law of gravitational attraction between them. Subsequent generations of mathematicians and physicists tried to extend Newton's analytical methods to the three-body problem (e.g., sun, earth, and moon) but curiously this problem turned out to be much more difficult to solve. After decades of effort, it was eventually realized that the three-body problem was essentially impossible to solve, in the sense of obtaining explicit formulas for the motions of the three bodies. At this point the situation seemed hopeless.

The breakthrough came with the work of Poincaré in the late 1800s. He introduced a new point of view that emphasized qualitative rather than quantitative questions. For example, instead of asking for the exact positions of the planets at all times, he asked: "Is the solar system stable forever, or will some planets eventually fly off to infinity?" [The latter may happen. See [2].] Poincaré developed a powerful *geometric* approach to analyzing such questions. That approach has flowered into the modern subject of dynamics, with applications reaching far beyond celestial mechanics. Poincaré was also the first person to glimpse the possibility of chaos, in which a deterministic system exhibits aperiodic behavior that depends sensitively on the initial conditions, thereby rendering long-term prediction impossible.

(C) The Origin of "Geometric" Topology. You can "do" analysis of singlevalued meromorphic functions in the plane, but "doing" analysis of multivalued meromorphic functions calls for finding suitable habitats for such functions. These habitats were introduced by Riemann—hence Riemann surfaces—in connection with his study of algebraic functions and their integrals, and are topologically spheres with handles. The number of handles of a Riemann surface is its genus. The preeminent role of this topological invariant of a Riemann surface in the characterization of its "meromorphic inhabitants" is described by Herman Weyl in a lengthy

passage in [3] which begins as follows:

> The classical example of the fruitfulness of the topological method is Riemann's theory of algebraic functions and their integrals. Viewed as a topological surface, a Riemann surface has just one characteristic, namely its connectivity number or genus p. For the sphere $p = 0$ and for the torus $p = 1$. How sensible it is to place topology ahead of function theory follows from the decisive role of the topological number p in function theory on a Riemann surface. I quote a few dazzling theorems: The number of linearly independent everywhere regular differentials on the surface is p. The total order (that is, the difference between the number of zeros and the number of poles) of a differential on the surface is $2p - 2$. If we prescribe more than p arbitrary points on the surface, then there exists just one single valued function on it that may have simple poles at these points but is otherwise regular; if the number of prescribed poles is exactly p, then, if the points are in general position, this is no longer true. The precise answer to this question is given by the Riemann-Roch theorem in which the Riemann surface enters only through the number p. If we consider all functions on the surface that are everywhere regular except for a single place P at which they have a pole, then its possible orders are all numbers $1, 2, 3, \ldots$ except for certain powers of p (the Weierstrass gap theorem). It is easy to give many more such examples. The genus p permeates the whole theory of functions on a Riemann surface. We encounter it at every step, and its role is direct, without complicated computations, understandable from its topological meaning (provided that we include, once and for all, the Thomson-Dirichlet principle as a fundamental function-theoretic principle).

In addition to his discovery and use of Riemann surfaces in function theory, Riemann—and independently of Riemann, Betti—discovered a class of topological invariants called Betti numbers. What are Betti numbers? Hilton and Pedersen [4] explain that

> With any topological space K we may associate certain abelian groups $H_r K$, $r = 0, 1, 2, \ldots$ called the homology groups of K, which, roughly speaking, count the r-dimensional 'holes' in K. If the space is n-dimensional then we only have homology groups up to dimension n. Then p_r is the *rank* of $H_r K$.

Betti numbers are intimately related to such topological household words as genus and Euler characteristic. Specifically, for a closed, orientable surface S_g the Betti numbers are $p_0 = 1$, $p_1 = 2g$, $p_2 = 1$. Since the Euler characteristic $\chi(S_g)$ of S_g is connected with its Betti numbers by the relation

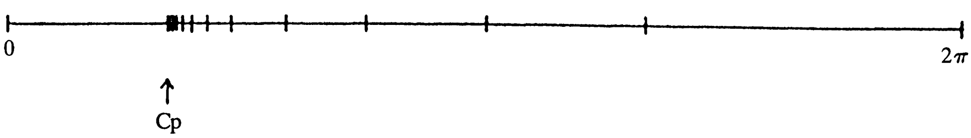
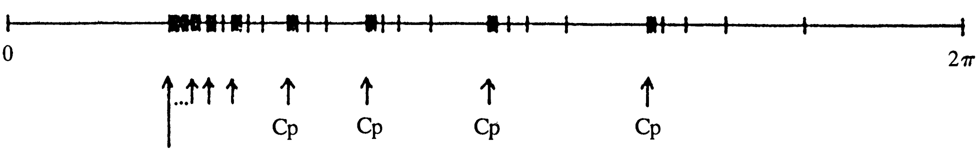
$$\chi = p_0 - p_1 + p_2,$$

it follows that

$$\chi(S_g) = 2 - 2g.$$

After Riemann, "geometric" topology was magnificently advanced at the end of the century by Poincaré, of whose topological contributions Dieudonné says (in his article on Poincaré in [11]) that "until the discovery of the higher homotopy groups in 1933, the development of algebraic topology was entirely based on Poincaré's ideas and techniques."

(D) The Origin of "General" Topology. "General," or point-set, topology came into being as a result of Cantor's interest in Fourier series. I quote from

> A... contribution of paramount importance to the emergence of point-set topology was... the work of Cantor. [An indirect indication of this is that the] dedication of Hausdorff's book [*Grundzuege der Mengenlehre*, 1914] reads: "To the creator of set theory *Georg Cantor* in grateful admiration"...
>
> Cantor had shown in 1870 that two Fourier series that converge pointwise to the same limit function have the same coefficients. In 1871 he improved this theorem by proving that the coefficients have to be the same also when convergence and equality of the limits hold for all points outside a finite exception set A in $[0, 2\pi]$. In a paper of 1872 he dealt with the problem of determining for which *infinite* exception sets uniqueness still holds. An infinite subset of $[0, 2\pi]$ must of course have at least one cluster point:
>
> 
>
>
> This is a very "innocent" example of an infinite subset of $[0, 2\pi]$. A somewhat "wilder" set is one whose cluster points themselves cluster around some point:
>
> 
>
>
> Cantor showed that if the sequence of subsets of $[0, 2\pi]$ defined inductively by $A^0 := A$ and $A^{n+1} := \{x \in [0, 2\pi] \| x$ is a cluster point of $A\}$ breaks off after finitely many terms, that is if eventually we have $A^k = \emptyset$, then uniqueness *does* hold with A as the exception set. ... [While] the motivation for Cantor's investigation stemmed from classical analysis and ultimately from physics, [it led him] to the discovery of a new type of subset A in $\mathbf{R}$ which must have been thought to be quite exotic, especially when the sequence $A, A^1, A^2, \ldots$ took a long time to break off. Now the subsets of $\mathbf{R}$ moved to the fore as objects to be studied for their own sake, and, what is more, studied from what we would recognize today as a topological viewpoint. Cantor continued along this path when later, while investigating general point sets in $\mathbf{R}$ and $\mathbf{R}^n$, he introduced the point-set topological approach on which Hausdorff could later base himself.

To flesh out the last sentence in this quote, let me add the following observation from [5] (p. 107):

> A not insignificant part of the nomenclature that is basic to general topology was introduced by Cantor in the course of his researches on point sets. Terms such as *perfect*, *dense*, *separable*..., *denumerable*, *continuous*, *closed*, *open*, are illustrative of the terms Cantor applied to point sets. Cantor obtained the essential theorems concerning the structure of these sets on the line, i.e. the topology of the line.

(E) From Geometry to Geometries. The first step in the transition from one geometry to many was the discovery of hyperbolic geometry. It became public knowledge in the late twenties and early thirties of the last century but it began to affect the thinking of large numbers of mathematicians only some forty years later. Its technical and intellectual significance for mathematics is immeasurable. One vital intellectual effect of its discovery was the realization, spelled out with all necessary precision by Gauss, that unlike arithmetic, the question of the geometry of physical space should be regarded as a question of physics.

In 1828 Gauss introduced an infinity of geometries, the so-called intrinsic geometries of surfaces in $\mathbf{R}^3$, in which the geodesics played the role of straight lines. This was a tremendous step forward from a mathematical viewpoint but not a conceptual revolution comparable to the discovery of hyperbolic geometry.

In his famous address of 1854 Riemann generalized Gauss' intrinsic geometry of a surface by extending it to n dimensions. Riemannian geometries are metric geometries.

In another famous address, in the Erlangen Program of 1872, Klein introduced his notion of a geometry as the totality of invariants of the subsets of a set acted upon by a group. Of the gap between these two conceptions Cartan had this to say [6]:

> The principle of general relativity brought into physics and philosophy the antagonism between the two leading principles of geometry due to Riemann and Klein respectively. The space-time manifold of classical mechanics and of the principle of special relativity is of the Klein type, and the one associated with the principle of general relativity is Riemannian. The fact that almost all phenomena studied by science for many centuries could be equally well explained from either viewpoint was very significant and persistently called for a synthesis that would unify the two antagonistic principles.

The geometric ideas of Riemann and Klein were combined in the concept of "spaces with connections." A space with a connection is a Riemannian space with a group of motions (Euclidean, affine, ...) grafted onto it. The group of motions is the "tool" needed to propel a vector parallel to itself along a curve in the Riemannian space.

Chronologically, the first step in this development was Levi-Civita's extension of the obvious notion of parallel transport of a vector along a curve in a plane to more general spaces. This was in 1917. In 1918, H. Weyl used a Riemannian space with a Euclidean connection in an attempt to unify electromagnetic and gravitational phenomena (see [6], [9], and [10]). "Weyl's ideas undoubtedly were the source from which E. Cartan, a few years later, developed his general theory of [spaces with] connections..." (see Dieudonne's article on H. Weyl in the Dictionary of Scientific Biography [11]).

A footnote on parallel transport. The remarks that follow give some indication of how one extends the notion of parallel transport from the plane to a more general space and mentions an unexpected application of this concept.

What it means to parallel-transport a vector along a straight line in the plane is suggested by Figure 1 below:

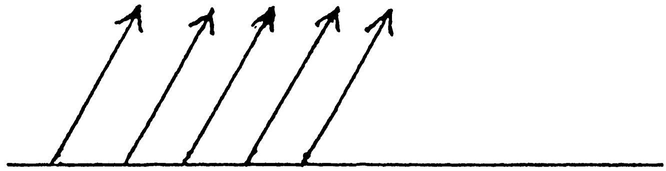


Figure 1

What is kept constant here is the angle between vector and straight line.

If we leave the plane for, say, a sphere and the straight line for a great circle G then, to obtain an analogue of Figure 1, we must draw tangent planes to the sphere at the points of G, tangent lines to G in the tangent planes, and draw vectors in the tangent planes that make the same angle with the appropriate tangent line.

It takes little to make this more general. "*Parallel translation of a vector tangent to a surface along a geodesic on this surface* is defined as follows: the point of origin of the vector moves along the geodesic, and the vector itself moves continuously so that its angle with the geodesic and its length remain constant."

It is obvious what to do if instead of a single geodesic arc we have a broken line consisting of several geodesic arcs. Finally, "*parallel translation of a vector along any smooth curve on a surface* is defined by a limiting procedure, in which the curve is approximated by broken lines consisting of geodesic arcs." (See Appendix 1: Riemannian [i.e. Gaussian] curvature, [7], pp. 301–317.)

What is the point of all this? Well, if you parallel-transport a vector around a triangle in a plane then, upon completion of the "tour,"you end up with a vector that coincides with the initial vector. But when you parallel-transport a vector around a spherical triangle—see Figure 2—with angles α, β, and γ then the end-vector makes an angle of $\alpha + \beta + \gamma - \pi$ with the initial vector. Now comes the surprise (see p. 21 in [9]): $\alpha + \beta + \gamma - \pi = R^{-2}S$, where R is the radius of

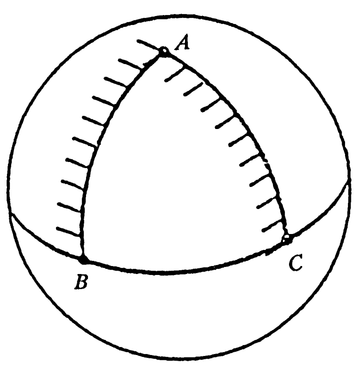


Figure 2

the sphere and S is the area of the triangle! In other words, *the angular change due to the parallel transport of a vector around a spherical triangle divided by its area is the (Gaussian) curvature of the sphere!* (If you are as use-obssessed as the writer of these lines, then you will enjoy the realization that this is, in principle, a prescription for finding out by means of a *local* test what kind of constant-curvature surface you are on!) Since anything related to the curvature of a manifold is important, parallel transport is important.

For more intuitive material on parallel transport see pp. 17–31 of **[8]**. For more precise material, see Ch. 8 of **[9]**.

(F) From Daniel Bernoulli to Distributions. Extrapolating from physical experiments, Daniel Bernoulli made the brilliant guess that the shape of a vibrating string is describable by means of a trigonometric series.

Somewhat later, Euler and d'Alembert arrived independently at essentially the same solution of the equation of the vibrating string which did not involve trigonometric series, and was thus very different from that of Bernoulli. While the two disagreed on the admissible initial shapes of the string, they agreed that Bernoulli's guess made no sense.

Shortly thereafter, Fourier's work, while sorely lacking in rigor, vindicated Daniel Bernoulli's informed guess and ushered in the study of Fourier series as a new branch of analysis. As for the resolution of the argument between Euler and d'Alembert, that had to await the introduction of distributions and of generalized solutions of differential equations. (The latter idea is found in Hilbert's rhetorical question at the end of his Problem 20: "Has not every regular variation problem a solution, ... provided also, if need be, that the notion of a solution shall be suitably extended?")

REFERENCES

1. Krzysztof Maurin, *Mathematics and Physics* in: Mathematical Lexicon (Polish). Wiedza Powszechna, 1993. Umemura's article is titled *Resolution of Algebraic Equations by Theta Constants*. This is an appendix in David Mumford's book *Tata Lectures on Theta II*, Progress in Mathematics, Vol. 43, Birkhaeuser, 1984.
2. D. G. Saari and Z. Xia, *Off to Infinity in Finite Time*. Notices of The AMS, May 1955, pp. 538–546.
3. H. Weyl, *Topology and Algebra as Two Roads of Mathematical Comprehension*. The Monthly, issues May and July–August, 1995.
4. P. Hilton and J. Pedersen, *The Euler Characteristic and Polya's Dream*.
5. J. H. Manheim, *The genesis of point set topology*. Pergamon Press, 1964.
6. E. Cartan, *Group Theory and Geometry* (French). L'Enseignement mathématique, 1927, pp. 200–225.
7. V. I. Arnold, *Mathematical methods of classical mechanics*. Springer, 1978.
8. N. Ya. Vilenkin, *In Search of Infinity*. Birkhaeuser (Boston), 1995.
9. J. A. Thorpe, *Elementary Topics in Differential Geometry*. Springer, 1979.
10. H. Weyl, *Space, Time, Matter*. Dover, 1922.
11. *Dictionary of Scientific Biography*. Published under the auspices of the American Council of Learned Societies, C. C. Gillispie, Editor in Chief. Charles Scribner's Sons, New York.

Department of Mathematics & Statistics
York University
North York, Ontario
CANADA M3J 1P3

THE AUTHORS

ALAN BEARDON obtained his Ph.D. at Imperial College, London in 1964 and then taught at the University of Maryland and the University of Kent at Canterbury. He is currently at the University of Cambridge where he has been since 1968.

His mathematical interests include complex analysis, Riemann surface and discrete groups, geometry (of all types), the iteration of analytic maps, and mathematical economics, and he has written four texts. His hobbies include making furniture, wood carving, collecting sea shells, and hill walking.

REINHOLD REMMERT was born in 1930 and received a Ph.D. in 1954 from the University of Münster in Westphalia. He taught mathematics at Münster, Ann Arbor, Erlangen, Johns Hopkins, Göttingen, Notre Dame (Indiana), Maryland, and Yale. He is one of the founders of the journal "Inventiones Mathematicae" and Chairman of the recently founded Society of Friends of Oberwolfach.

LOUIS ZULLI received his B.S. from SUNY Stony Brook and his Ph.D. from Cornell University, where his advisor was Allen Hatcher. He is currently an Evans Instructor at Rice University in Houston. A low-dimensional topologist by training, Dr. Zulli has done research in knot theory and 3-manifold topology. He is also interested in popularizing mathematics, and enjoys the special challenge of teaching and writing about the subject. His hobbies include softball, hiking, chess, and mycology.

MARGARET KLEINFELD received a B.A. in physics from the University of Rochester in 1960 and a Ph.D. in mathematics from Syracuse University in 1965, where she did her thesis under the direction of Erwin Kleinfeld whom she later married. Margaret and Erwin have been on the faculty of the University of Iowa since 1968. They spent 1985–86 as consultants at the University of Indonesia in Jakarta and during 1988–89 they taught in a MUCIA program in Shah Alam, Malaysia. They pursue research in non associative algebra and have recently authored two textbooks on linear algebra that will appear soon from Nova Science Publishing.

DORON ZEILBERGER wrote so far 6 shorter notes and one main article for the Monthly. 100% of his main articles received the prestigious Lester R. Ford award. If the article in this issue does not receive it, it would drop to 50%, which is still respectable. His specialty is combinatorics, and his hobbies include number theory, logic, and algebra. His only fault is his quick temper. He also enjoys playing with his pet computer, Shalosh B. Ekhad.

THOMAS HULL received his B.A. from Hampshire College and the M.S. in mathematics from the University of Rhode Island, where he now pursues the Ph.D., specializing in graph theory. He has been folding paper since he was eight years old and co-authored the book *Origami, Plain and Simple* with American origami master Robert Neale. He has lectured on the mathematics of origami around the world, including at the recent 2nd International Origami Science Meeting in Otsu, Japan. He hopes to spread the word of origami math to the mathematics community by writing a book on the subject. *Origametry*. But this thing called a thesis keeps getting in the way... .

GARY R. LAWLOR was an undergraduate at Brigham Young University and a graduate student at M.I.T. and Stanford. His first job began in 1988 as an N.S.F. postdoctoral fellow at Princeton University, where he was later promoted to assistant professor. He returned in 1991 to BYU and the mountain valleys where his love of mathematics had first sprung up—the gift of a dedicated dad who made math fun (from times table tricks to a googol plex), and who made him believe he could do anything in the world.

VICTOR J. KATZ spent the 1994–95 academic year as Visiting Mathematician at the Mathematical Association of America, although he was only visiting from the University of the District of Columbia, three subway stops away, where he is Professor of Mathematics. His major interest is in the history of mathematics. His recent textbook, *A History of Mathematics: An Introduction* (Harper Collins, 1993), was awarded the 1995 Watson Davis and Helen Miles Davis Prize by the History of Science Society for an outstanding introductory book in some aspect of the history of science.

If A_n Has $6n$ Dyes in a Box, With Which He Has To Fling [at least] n Sixes, Then A_n Has An Easier Task Than A_{n+1}, at Eaven Luck

The probability of A_n succeeding, $1 = \sum_{k=0}^{n-1} \binom{6n}{k} (\frac{1}{6})^k (\frac{5}{6})^{6n-k}$, can be expressed in a form that makes the asserted monotonicity evident:

$$1 - \sum_{m=0}^{n-1} \frac{2(94500m^4 + 214830m^3 + 171573m^2 + 56243m + 6250)(6m)!5^{5m+2}}{(5m+5)!m!6^{6m+5}}$$

The cases $n = 1$ and 2 were the subject of a fascinating sequence of letters between Isaac Newton and Samuel Pepys (1693) that is reproduced in *The American Statistician* 3 (1960), 27–30.

Contributed by Doron Zeilberger, Temple University, who derived the displayed expression using `zeillim` in the package EKHAD (`http://www.math.temple.edu/~zeilberg`) that accompanies M. Petkovsek, H. S. Wilf, and D. Zeilberger, *A = B*, A. K. Peters, Wellesley, 1996.

NOTES

(**10510**) This disproves the conjecture of Allan B. Calhamer, "The only cube which equals a square plus a triangular number is 64.", quoted on p. 240 of Underwood Dudley, *Mathematical Cranks*, MAA Spectrum, 1992. (**10512**) Notation and conventions of the theory of convex sets used in the problem are: $\mathrm{Conv}(Q)$ for the *convex hull* of Q, which is the smallest convex set containing Q; and $\mathrm{Len}(\partial\Phi)$ for the length of the perimeter of the convex set Φ. In the special case in which Φ degenerates to a line segment of length l, $\mathrm{Len}(\partial\Phi) = 2l$, since one must traverse the segment twice in order to have the effect of going *around* it. (**10513**) Here, $\overline{A}$ and $\overline{B}$ denote the (entrywise) complex conjugates of A and B, respectively. (**10514**) All subscripts are taken modulo 6, so that $P_7 = P_1$, $Q_0 = Q_6$, etc. in the description of the construction. The construction of the outward equilateral triangles is a generalization of that used in Napoleon's Theorem. See J. E. Wetzel, "Converses of Napoleon's Theorem", this MONTHLY, 99 (1992), 339–351, for information on that topic. A related problem by the same author appears in *Mathematics Magazine* as problem 1493 (February 1996). In that problem, the properties sought here for the hexagon $R_1R_2\ldots R_6$ are considered for what is denoted here by $Q_1Q_2\ldots Q_6$. Although the notation is different, the figure included with the statement in the *Magazine* may be useful for this problem as well.

SOLUTIONS

Comparing Unlabeled Graphs by Parity

10285 [1993, 185]. *Proposed by Frank Schmidt, Arlington, VA.*

Let e_n, respectively o_n, denote the number of unlabeled graphs on n vertices having an *even*, respectively *odd*, number of edges. Show that $e_n \geq o_n$ for all n.

Solution I by Robin J. Chapman, University of Exeter, Exeter, U. K.. We use Burnside's Lemma: If X is a finite set, and G is a finite group acting on X, then the number of orbits in X under G is $\frac{1}{|G|}\sum_{g\in G}|X_g|$, where $X_g = \{x \in X : gx = x\}$. Partition the (labeled) graphs with vertex set $[n] = \{1,\ldots,n\}$ into sets E, O by parity of the number of edges. Let G be the symmetric group S_n, and let the action of G partition E and O into e_n and o_n orbits, respectively. Hence it suffices to prove that $|E_g| \geq |O_g|$ for all $g \in S_n$.

A graph is fixed by $g \in S_n$ if and only if its edge set is the union of orbits of edges of K_n induced by applying g to $[n]$. If these orbits all have even length, then O_g is empty and trivially $|E_g| \geq |O_g|$. If there is an orbit A of odd length, then the operation of symmetric difference with A establishes a bijection between E_g and O_g, yielding $|E_g| = |O_g|$. This completes the proof.

If K_n has an odd number of edges (when $n \equiv 2$ or $3 \pmod 4$), then complementation establishes a bijection between E and O, and $e_n = o_n$. Otherwise, S_n has an element g with $\lfloor n/4 \rfloor$ 4-cycles, and all the orbits of edges under g have even length. In this case $|E_g| > |O_g|$ and $e_n > o_n$.

Solution II by Valery A. Liskovets, Minsk, Belarus. Let $f_n(x)$ be the ordinary generating function for unlabeled n-vertex graphs by number of edges. Setting $x = -1$ subtracts those

of odd size from those of even size, computing $e_n - o_n$. By Pólya's Theorem, the generating function $f_n(x)$ is obtained by letting $x_k = 1 + x^k$ for each k in the cycle index for the pair group $S_n^{(2)}$, which is the permutation group on the edges $E(K_n)$ induced by applying S_n to the vertices. Denoting the cycle index by $Z(S_n^{(2)}; x_1, x_2, \ldots)$, we obtain $e_n - o_n = Z(S_n^{(2)}; 0, 2, 0, 2, \ldots)$. Since $Z(S_n^{(2)})$ is a polynomial with non-negative coefficients, this value is non-negative. As observed by F. Harary and E. M. Palmer *Graphical Enumeration*, Academic Press, 1973, formula 6.2.3, the result of this computation also equals the number of self-complementary unlabeled n-vertex graphs (and hence $e_n = o_n$ unless $n \equiv 0$ or 1 (mod 4)). The technique yields analogous results for digraphs and other objects.

Solved also by B. Doran, I. Kastanas, J. Kuplinski, K.-W. Lau (Hong Kong), C. R. Pranesachar (India), A. N. 't Woord (The Netherlands), and the proposer. One incorrect solution was received.

Polynomials with Small Positive Roots

10286 [1993, 185]. *Proposed by Călin Popescu, Université Catholique de Louvain, Louvain, Belgium.*

Let $a_0 + a_1x + \ldots + a_mx^m$ be a polynomial with real coefficients and $(-1)^m a_m > 0$. Suppose that all roots of this polynomial are positive real numbers less than 1. Prove that

$$(-1)^{p+1} \sum_{k=n}^{m-p-1} (-1)^k \binom{k}{n} \sum_{h=0}^{m} a_h \sum_{i+j=k} (-1)^j \binom{h}{i} \binom{m-h}{j} > 0$$

for all nonnegative integers n and p whose sum is less than m.

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K.. Let

$$f(x) = \sum_{h=0}^{m} a_h x^h,$$

and let $S_{n,p}$ denote the left side of the desired inequality. We have

$$S_{n,p} = (-1)^{p+1} \sum_{k=n}^{m-p-1} \binom{k}{n} b_k,$$

where

$$b_k = \sum_{h=0}^{m} a_h \sum_{i+j=k} (-1)^i \binom{h}{i} \binom{m-h}{j}$$

is the coefficient of x^k in

$$g(x) = \sum_{h=0}^{m} a_h (1-x)^h (1+x)^{m-h} = (1+x)^m f\left(\frac{1-x}{1+x}\right).$$

Multiplying f and g by a positive constant does not change the roots or the sign of $S_{n,p}$. The coefficients of a polynomial with positive roots alternate in sign. Hence we have $a_0 > 0$ and may divide by a_0 to assume $a_0 = 1$. Now we have $f(x) = \prod_{j=1}^{m}(1 - \alpha_j x)$ with each $\alpha_j > 1$, and we compute

$$g(x) = \prod_{j=1}^{m}[(1+x) - \alpha_j(1-x)] = b_0 \prod_{j=1}^{m}(1 - \beta_j x),$$

where $b_0 = \prod_{j=1}^{m}(1 - \alpha_j) \neq 0$ and $\beta_j = (\alpha_j + 1)/(\alpha_j - 1) > 1$. Hence the roots of $g(x)$ lie in $(0, 1)$. By Rolle's Theorem, the roots of each derivative of g also lie in $(0, 1)$.

Let h be the nth derivative of g. If we write $h(x) = \sum_{j=0}^{m-n} c_j x^j$, then $c_j = n!\binom{j+n}{n} b_{j+n}$, and we have $S_{n,p} = \frac{(-1)^{p+1}}{n!} \sum_{k=0}^{m-(n+p+1)} c_k$. Hence we must show that $\sum_{k=0}^{t} c_k$ has the same sign as $(-1)^{m-n-t}$ for all t with $0 \le t \le m-n$.

First suppose $t = 0$. Since c_0 has the same sign as b_n, the alternation of signs implies that this has the same sign as $(-1)^n b_0$, which equals $(-1)^n \sum_{h=0}^{m} a_h$. Since $\sum_{h=0}^{m} a_h = f(1)$, which is above all roots, $f(1)$ has the same sign as a_m. Hence c_0 has the same sign as $(-1)^{m-n}$.

We now renormalize to assume $c_0 = 1$, without changing the roots of h. It suffices to show

Lemma. *If $c_0 = 1$ and $h(x) = \sum_{i=0}^{r} c_i x^i$ has r roots in $(0, 1)$, then the sign of $\sum_{i=0}^{t} c_i$ is $(-1)^t$ for each t with $0 \le t \le r$.*

Proof. We prove the result by induction on r, taking the easily verified case $r = 0$ as basis.

For $r > 0$, consider first $t = r$. In this case, the sum equals $h(1) = \prod_{j=1}^{r}(1 - \gamma_j)$, where $\gamma_j > 1$ are the reciprocals of the roots, and hence the sign is $(-1)^r$.

Suppose now that $r > t$. Isolating one of the factors, we have $h(x) = (1 - \gamma x)h_1(x)$, where $\gamma > 1$ and $h_1(x) = \sum_{i=0}^{r-1} d_i x^i$ is a polynomial of degree $r - 1$ with roots in $(0, 1)$. For $i > 0$, $c_i = d_i - \gamma d_{i-1}$. Hence

$$1 + c_1 + \ldots + c_t = (1 + d_1 + \ldots + d_t) - \gamma(1 + d_1 + \ldots + d_{t-1}).$$

By the induction hypothesis, both contributions have the same sign as $(-1)^t$, which completes the proof.

Editorial comment. For the final argument, Antonios D. Melas normalized the polynomial to have $c_r = 1$, and noticed that the coefficients c_j are linear in each root. Thus, the statement follows from the special case in which all roots take the extremal values of 0 or 1, which may be verified directly.

Solved also by O.P. Lossers and A.D. Melas (Greece).

Telescoping Series of Arctangents

10292 [1993, 291]. *Proposed by Jean Anglesio, Garches, France.*

Obtain explicit values for the following series.

(a) $\displaystyle\sum_{n=1}^{\infty} \arctan\left(\frac{2}{n^2}\right)$

(b) $\displaystyle\sum_{n=1}^{\infty} \arctan\left(\frac{8n}{n^4 - 2n^2 + 5}\right)$

Composite solution I by many readers. The answers are, respectively, $3\pi/4$ and $\pi - \arctan(1/2)$. Since

$$\arctan\frac{2}{n^2} = \arctan(n+1) - \arctan(n-1)$$

and

$$\arctan\frac{8n}{n^4 - 2n^2 + 5} = \arctan\frac{(n+1)^2}{2} - \arctan\frac{(n-1)^2}{2},$$

the sum for even n and that for odd n each telescope. The terms for n odd give $\pi/2$ in each case, while those for n even give $\pi/4$ in **(a)** and $\pi/2 - \arctan(1/2)$ in **(b)**.

Solution II by Donald A. Darling, Newport Beach, CA. There is a systematic method for treating similar problems. Let $P(x)/Q(x)$ be a rational function with real coefficients in

which $\deg Q \geq 2 + \deg P$, and assume for simplicity that the leading coefficient of $Q(x)$ is 1. On using the well-known relation

$$\arctan y = \frac{1}{2i} \log \frac{1+iy}{1-iy}$$

for y real, we obtain

$$\begin{aligned}\sum_{n=1}^{\infty} \arctan \frac{P(n)}{Q(n)} &= \frac{1}{2i} \sum_{n=1}^{\infty} \log \frac{Q(n)+iP(n)}{Q(n)-iP(n)} \\ &= \frac{1}{2i} \log \prod_{n=1}^{\infty} \frac{Q(n)+iP(n)}{Q(n)-iP(n)}.\end{aligned}$$

Let $a_1, a_2, \ldots, a_k$ be the zeros of $Q(x) - iP(x)$. Since $P(x)$ and $Q(x)$ have real coefficients, the zeros of $Q(x) + iP(x)$ are the complex conjugates of the a_j. Thus

$$\begin{aligned}\sum_{n=1}^{\infty} \arctan \frac{P(n)}{Q(n)} &= \frac{1}{2i} \log \prod_{n=1}^{\infty} \frac{(n-\overline{a_1})(n-\overline{a_2})\cdots(n-\overline{a_k})}{(n-a_1)(n-a_2)\cdots(n-a_k)} \\ &= \frac{1}{2i} \log \frac{\Gamma(1-a_1)\Gamma(1-a_2)\cdots\Gamma(1-a_k)}{\Gamma(1-\overline{a_1})\Gamma(1-\overline{a_2})\cdots\Gamma(1-\overline{a_k})}\end{aligned}$$

where we have used the formula of section 12.13 of E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Fourth Edition, Cambridge, 1927. Thus,

$$\sum_{n=1}^{\infty} \arctan \frac{P(n)}{Q(n)} = \sum_{j=1}^{k} \arg\left(\Gamma(1-a_j)\right)$$

gives the desired value modulo 2π. The separation of the possible values allows simple estimates to isolate the correct value.

In (**a**), $P(x) = 2$ and $Q(x) = x^2$, so the a_j are the zeros of $x^2 - 2i$, which are $1+i$ and $-1-i$. The formula gives

$$\begin{aligned}\arg\left(\Gamma(-i)\right) + \arg\left(\Gamma(2+i)\right) &= \arg\left(\Gamma(-i)\right) + \arg(1+i) + arg(i) + \arg\left(\Gamma(i)\right) \\ &= \arg(1+i) + arg(i) = \frac{\pi}{4} + \frac{\pi}{2} = \frac{3\pi}{4}.\end{aligned}$$

In (**b**), $Q(x) - iP(x) = x^4 - 2x^2 - 8ix + 5 = (x^2+1)^2 - 4(x+i)^2$, so the roots are $-2+i$, $-i$ (with multiplicity 2), and $2+i$, and the formula gives

$$\arg(2-i) + \arg(1-i) + \arg(i) + \arg(-1+i) = \pi - \arctan\left(1/2\right).$$

In both cases, the sum of the first few terms shows that no modification by a multiple of 2π is needed.

Editorial comment. Several readers used the approach of Solution I giving varying amounts of detail to establish the formulas. Typically, as in a joint solution by Sammy Yu (age 13 at time of writing) and Jimmy Yu (age 11 at time of writing) the usual formula for simplifying $\tan(\alpha - \beta)$ was used.

Erhard Braune formulated

Theorem. *Let L and a be a positive integers. Define $p(n) = \sum_{j=0}^{L-1} (n+a)^j (n-a)^{L-1-j}$ and $q(n) = (n+a)^L (n-a)^L$. Then for any $K > 0$,*

$$\sum_{n=a}^{\infty} \arctan \frac{Kp(n)}{q(n)+K^2} = \frac{\pi}{2} + \sum_{j=1}^{2a-1} \arctan \frac{K}{j^2}.$$

Solution II shows how one might *discover* the answer. In particular, this also applies to $\sum_{n=1}^{\infty}(\arctan)1/n^2$, considered in E 3375 [1990, 239; 1991, 652]. H. K. Krishnapriyan mentioned Allen R. Miller and H. M. Srivastava, "On Glaisher's infinite sums involving the inverse tangent function", *Fibonacci Quart.* 30 (1992), 290–294, which traced a generalization of (**a**) to a paper of J. W. L, Glaisher in 1878, mentioned other appearances of these sums as formulas in tables and as problems, and gave some new results. Proofs used the Euler-Maclaurin summation formula

$$\sum_{k=0}^{n} f(k) = \int_0^n + \frac{1}{2} f(0) + \frac{1}{2} f(n) + \int_0^n P(x) f'(x)\, dx$$

where $P(x) = x - \lfloor x \rfloor - 1/2$.

This formula is useful for keeping track of the branch of the complex logarithm. It was not explicitly used in submitted solutions, although other relations between sums and integrals did appear.

Solved by 57 readers (including those cited) and the proposer. One incorrect solution was received.

Polynomialrecurrençology

10300 [1993, 401]. *Proposed by Eric H. Mason, McDonnell Douglas Corporation, McLean, VA.*

Let $\{P_m(z) : m = 1, 2, 3, \ldots\}$ be the sequence of polynomials defined by $P_1(z) = z - 1$, $P_2(z) = z^2 - z - 1$, and $P_m(z) = z P_{m-1}(z) - P_{m-2}(z)$ for $m > 2$. Show that the roots of $P_m(z)$ are $2\cos\big((2k-1)\pi/(2m+1)\big)$ for $1 \le k \le m$.

Solution I by Stephen M. Gagola, Jr., Kent State University, Kent, OH. Notice that $P_1(z + z^{-1}) = z - 1 + z^{-1}$ and $P_2(z + z^{-1}) = (z + z^{-1})^2 - (z + z^{-1}) - 1 = z^2 - z + 1 - z^{-1} + z^{-2}$. The recurrence relation for $P_n(z)$ now easily implies that $P_m(z + z^{-1}) = z^m - z^{m-1} + \cdots + z^{-m}$ for all m. Thus

$$P_m(z + z^{-1}) = z^{-m}\Big(z^{2m} - z^{2m-1} + \cdots + 1\Big) = z^{-m} \frac{z^{2m+1} + 1}{z + 1}.$$

Substituting $z = e^{(2k-1)\pi i/(2m+1)}$ and using $z^{2m+1} = -1$ shows that $z + z^{-1} = 2\cos\frac{(2k-1)\pi}{2m+1}$ is a root of P_m for $1 \le k \le m$. As P_m has degree m, we have found all the roots of P_m.

Solution II by John Todd, California Institute of Technology, Pasadena, CA. This seems a routine problem to be attacked by a generating function

$$g(t) = \sum_{m=1}^{\infty} P_m(z) t^{m-1}.$$

The initial conditions give

$$g(t)\big(1 - zt + t^2\big) = z - 1 - zt$$

Writing $z = 2\cos\theta$, the partial fraction representation of $g(t)$ is

$$g(t) = A\big(1 - te^{i\theta}\big)^{-1} + B\big(1 - te^{-i\theta}\big)^{-1},$$

where

$$A = \frac{e^{2i\theta}}{1 + e^{i\theta}} \text{ and } B = \frac{e^{-i\theta}}{1 + e^{i\theta}}.$$

Expanding, we find

$$\begin{aligned} P_m(z) &= Ae^{(m-1)i\theta} + Be^{-(m-1)i\theta} \\ &= Be^{-(m-1)i\theta}\Big((A/B)e^{2(m-1)i\theta} + 1\Big). \end{aligned}$$

For this to vanish, we must have

$$e^{(2m+1)i\theta} = -1.$$

Hence $\theta = (2k-1)\pi/(2m+1)$, $k = 1, 2, \ldots, m$, as required.

Solution III by Theodore J. Rivlin, Thomas J. Watson Research Center, Yorktown Heights, NY. $U_n(x)$, the Chebyshev polynomial of the second kind, satisfies

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}.$$

Hence, $U_{-1}(x) = 0$, $U_0(x) = 1$, and $U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x)$, $n = 0, 1, \ldots$. Let $S_n(x) = U_n(x/2)$, $n = -1, 0, 1, \ldots$. We claim that $P_m(x) = S_m(x) - S_{m-1}(x)$, $m = 1, 2, \ldots$. This is verified directly for $m = 1$ and $m = 2$, and subsequent cases follow by induction since these expressions satisfy the same recurrence. Now,

$$P_m(2\cos\theta) = \frac{\cos(2m+1)\theta/2}{\cos\theta/2},$$

and hence the zeros of $P_m(x)$ are as claimed.

Solution IV by David Callan, University of Wisconsin, Madison, WI. Let $A_n = \big(a_{ij}\big)$ denote the tridiagonal matrix with 1's on the super and infra diagonals (where $|i-j| = 1$), $a_{nn} = 1$, and 0's elsewhere. Expanding $\det(\lambda I_n - A_n)$ by cofactors, it is easy to see that it satisfies the same recurrence relation and initial conditions as $P_n(\lambda)$; hence $P_n(\lambda) = \det(\lambda I_n - A_n)$, and the roots of P_n are the eigenvalues of A_n. Now, it is routine, if a little tedious, to verify that the eigenvalue-eigenvector pairs $\lambda_k, \mathbf{u}_k$, $k = 1, \ldots, n$, for A_n are given by

$$\lambda_k = 2\cos\frac{(2k-1)\pi}{2n+1}, \quad \mathbf{u}_k^{\mathrm{T}} = \left(\sin\frac{j(2k-1)\pi}{2n+1}\right)_{j=1}^{n}.$$

Editorial comment. If we regard the use of the characteristic polynomial of the recurrence as being a variant on solution II, the 53 solutions (David Callan and Costas Efthimiou each submitted two solutions) may be classified as follows: 14 most resembled solution I; 16 most resembled solution II; 21 most resembled solution III; 2 most resembled solution IV.

The connection with Chebyshev polynomials suggests that these may be orthogonal polynomials. The Jacobi polynomial formula

$$P_n^{(-1/2,1/2)}(\cos\theta) = \frac{(2n)!}{2^{2n}(n!)^2}\frac{\cos(2n+1)\theta/2}{\cos\theta/2}$$

shows that this is the case. Jacobi polynomials are orthogonal on $[-1, 1]$ with weight $(1-x)^{-1/2}(1+x)^{1/2}$.

Solution IV also applies to the matrix in which $a_{nn} = 0$, along with the rest of the main diagonal. H. K. Krishnapriyan noted that this simpler matrix is covered by E 1678 [1964, 317; 1965, 190]. The solution to that problem includes many references that are also relevant here.

Solved by 50 readers (including those cited) and the proposer.

Prime Real Algebraic Integers

10305 [1993, 402]. *Proposed by J.C. Lagarias, AT&T Bell Laboratories, Murray Hill NJ.*

Is there a smallest prime number if the set of primes is enlarged to include certain real algebraic numbers? In particular:

(**a**) Call a real algebraic number α an A-prime number if it is a totally real algebraic integer such that

(i) The ideal (α) is a prime ideal in the ring of integers of the field $\mathbb{Q}(\alpha)$.

(ii) $\alpha \geq |\sigma(\alpha)|$ for all algebraic conjugates $\sigma(\alpha)$ of α.

Is there a smallest A-prime number?

(**b**) Call a number an A^*-prime number if it is a real algebraic integer (not necessarily totally real) satisfying (i) and (ii) above.

Is there a smallest A^*-prime number?

Solution by Ilias Kastanas, California State University, Los Angeles, CA. For (**a**), we prove that every totally real algebraic integer whose conjugates all have absolute value less than $\sqrt{2}$ is in the set $\{-1, 0, +1\}$. Since $\sqrt{2}$ is A-prime, this implies that $\sqrt{2}$ is the smallest A-prime.

Suppose α is a totally real algebraic integer whose conjugates have absolute value less than $\sqrt{2}$. Let $f(x) = \prod(x - \alpha_i)$ be the minimum polynomial of α, and put $g(x) = \prod(x - \alpha_i^2)$. The coefficients of $g(x)$ are algebraic integers invariant under all automorphisms of the splitting field, so they are natural integers. (Alternatively, we can construct $g(x)$ directly in $\mathbb{Z}[x]$ from $h(x) = (-1)^n f(x) f(-x)$. Since $h(x) = h(-x)$, all its odd-order derivatives vanish at 0 and $h(x)$ is a polynomial in x^2. With $h(x) = g(x^2)$, the roots of $g(x) = 0$ are the α_i^2.) Now $j(x) = g(x+1)$ has roots $\alpha_i^2 - 1$. If $\alpha \neq 0$, then $0 < \alpha_i^2 < 2$, so $\left|\alpha_i^2 - 1\right| < 1$. Hence $|h(0)| = \left|\prod(\alpha_i^2 - 1)\right| < 1$. But $h(0) \in \mathbb{Z}$, so $h(0) = 0$, meaning that some α_i^2 is 1.

For (**b**), the nth root of 2 has norm ± 2, for integer $n \geq 2$, so the integer roots of 2 are A^*-prime. Thus a smallest A^* prime would have absolute value at most 1. In view of (ii), its norm would be ± 1, making it a unit, which is a contradiction.

Solved also by H. von Eitzen (Germany), A. N. 't Woord (The Netherlands), and the proposer.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

The Lighter Side of Mathematics. Edited by Richard K. Guy and Robert E. Woodrow, The Mathematical Association of America, Inc., 1994; viii + 367, \$38.50.

Reviewed by **Victor J. Katz**

Recreational mathematics has a history nearly as long as mathematics itself. In fact, some of the earliest mathematical documents have problems which can well be regarded as "recreational." For example, the Rhind Mathematical Papyrus, dating from about 1650 B.C. in Egypt, contains a problem which simply lists: houses, 7; cats, 49; mice, 343; spelt, 2401; *hekat* (a measure of grain), 16807, along with the total, 19607 [**3**, p. 59]. (The author also finds the sum of this geometric series in a more efficient way.) The general assumption is made that the problem is intended to be a puzzle problem: In each of 7 houses are 7 cats; each cat kills 7 mice, each mouse would have eaten 7 ears of spelt, and each ear of spelt would have produced 7 *hekat* of grain. How much grain is thereby saved? It may well be that this is a forerunner of the well known, "As I was going to St. Ives, ... " And the Chinese *Nine Chapters on the Mathematical Art*, compiled around the beginning of our era, has the earliest known example of a problem familiar to everyone: There is a reservoir with five channels bringing in water. If only the first channel is open, the reservoir can be filled in 1/3 of a day. The second channel by itself will fill the reservoir in 1 day, the third channel in 2 1/2 days, the fourth one in 3 days, and the fifth one in 5 days. How long will it take to fill the reservoir if all the channels are open together? [**6**, p. 68] This problem in one version or other shows up in virtually every algebra text since that time.

It is difficult perhaps to make a distinction between "recreational" and "serious" mathematics. After all, much of mathematics was developed by people having fun with certain ideas. Often, the germ of the idea came from a "practical" problem, but after that problem was solved, the mathematician simply followed out all sorts of consequences of the solution, well beyond any immediate practical value. The problems above are examples of those with no practical use, although they probably grew out of real situations.

Since similar recreational problems occur in various civilizations at different times, it is possible that it was those problems which were most easily transmitted from one culture to another. Among other such problems which have traveled widely are the hundred fowls problem, the river crossing problem, and the construction of magic squares. The first appearance of the former is in fifth century China, where Zhang Quijian posed it: A rooster is worth 5 coins, a hen 3 coins, and 3 chicks 1 coin. With 100 coins we buy 100 of the fowls. How many roosters, hens, and chicks are there? [**5**, p. 277] The problem reappeared over the

centuries in various guises in mathematical problem collections in India, the Islamic world, and Europe. The second problem, in which a man must ferry three items across a river only one at a time, is also widespread. [**1**, p. 109 ff.] The idea, of course, is that certain items may not be left together unattended. (In the first known version of the problem, the items were a wolf, a goat, and a head of cabbage.) This problem and its numerous variations, some requiring quite a large number of river crossings, recur not only throughout Western and Eastern Europe, but also in many African countries and in African-American folklore. Magic squares seem to have their origin in China, with legends taking them back to the third millennium B.C. The problem of construction of such squares was extensively studied in medieval Islam, with one of the most detailed discussions of these constructions being due to Muhammad ibn Muhammad, a Nigerian who lived in the early 18th century. [**7**, pp. 137–151] Each of these problems, whatever their origin, has led to interesting and serious mathematical ideas. They have also been vehicles for making mathematics appealing.

With recreational mathematics having such a wide-ranging history, it is not surprising that books and articles dealing with the subject have always been popular. The columns of Martin Gardner are the best known modern examples of this kind of mathematical writing, but other famous writers in this genre over the past century include Lewis Carroll, Henry Ernest Dudeney, and Sam Loyd. The classic book in this field in the English language is W. W. Rouse Ball's *Mathematical Recreations & Essays*, which has been in print, in numerous editions, since it first appeared in 1892. [**2**] Ball's work discusses problems of all sorts, including arithmetical and geometrical problems and games, polyhedra and tessellations, chess-board problems, magic squares, coloring problems, graph and network problems, and various "unclassifiable" problems. These types of problems are what are classically called "recreational," mostly because the statements of the problems, if not the solutions, can be understood by virtually anyone.

But what is the status of "recreational" mathematics today? Do "real" mathematicians do "recreational mathematics?" If we look at the programs of recent major international mathematical meetings, we see some kind of answer to these questions. Namely, the International Congresses of Mathematicians rarely, if ever, have aspects of their program devoted to such topics, although, of course, there are topics considered at the Congresses which have emerged from recreational activities. On the other hand, the International Congresses of Mathematics Education in recent years have always included sections devoted to Mathematical Games and Recreations, and, indeed, have one scheduled for the 1996 Congress in Seville. Evidently, mathematics educators, if not mathematicians, continue to believe in the power of mathematical games and recreations to inspire and attract students into mathematics.

In recent years, there have also been several major meetings entirely devoted to the topic of recreational mathematics. The book under review is the proceedings of one of them, the Eugène Strens Memorial Conference on Recreational Mathematics and its History, held at the University of Calgary in 1986. As we learn from the opening article of the book, Eugène Strens (1899–1980) was a Dutch engineer who, among many other interests, collected an immense library of books on recreational mathematics. A few years after Strens' death, Professor Guy arranged for the collection to be transferred to the University of Calgary Library, where it now forms the heart of a growing special collection of materials on recreational mathematics. It was thus appropriate that this international meeting, attended by about one hundred participants from a dozen countries, took place in Calgary.

The authors represented in the book provide another answer to the question of whether real mathematicians do recreational mathematics, because many of them are "serious" mathematicians. Many of their contributions are quite serious mathematical articles, complete with definitions, theorems, discussions, and proofs. It is not entirely clear why some of these articles belong in a book devoted to "the lighter side" of mathematics, whatever that may mean. In fact, the variety of articles included makes it difficult to determine what the audience for this book ought to be. A student enamored of the type of recreational mathematics included in Ball's work would not even be able to understand the problem in many of the articles, let alone the solution, because these often require a strong mathematical background. Evidently, the editors did not insist on any uniformity from the authors, and the authors responded with articles ranging from a biographical sketch of one of the more famous puzzle creators to detailed mathematical papers already published in respectable journals of research mathematics. And, as is unfortunately often the case with volumes of this sort, some of the articles read like transcripts of the presentations given at the meeting, complete with crucial references to actual displays, such as mathematical games and toys, which could not be included in this printed version. Readers intrigued by some of the references will have to search catalogs for the actual toys so they can have the same experiences that conference participants had.

Despite its unevenness, however, the book contains many gems, including prize-winning studies from other publications and summaries of important ideas by excellent expository writers. Some of the contributions are biographical or discuss the history of certain aspects of recreational mathematics and may well provide ideas for further research in the relationship of recreational to "serious" mathematics. But the rest of this review will provide a brief guide to the best mathematical papers in *The Lighter Side*, most of which deal with topics classically called "recreational."

A name which appears often in *The Lighter Side* is that of Maurits Escher, and we are treated to a superb introduction to his mathematical analysis of tilings by Doris Schattschneider, the MAA's Hedrick lecturer for 1995. Schattschneider, in "Escher: A Mathematician in Spite of Himself," shows how Escher, who denied understanding mathematics, in fact created for himself a set of categories, invented a classification scheme, and explored various questions of colored periodic tilings of the plane in a systematic fashion. We even see some of his notebook pages which describe transitions from one type of tiling to another, mathematical descriptions which are then mirrored in some of his graphical works.

In another article about tiling, Stan Wagon shows us "Fourteen Proofs of a Result About Tiling a Rectangle." The result in question asserts that whenever a rectangle is tiled by rectangles each of which has at least one integer side, then the tiled rectangle has at least one integer side. It turns out that there are numerous ways of approaching this theorem, each of which has its own appeal, and some of which are easily generalizable to more complex situations.

Daniel Ullman discusses "The Road Coloring Problem" to show how a tourist bureau can mark its road intersections to be able to guide lost tourists to their destination by a simple series of instructions which is not dependent on where the tourist is. Turning this "recreational" problem into a graph theoretical problem, the author describes what is known and what remains to be learned about this puzzle.

Have you ever wondered how your bowling score compares with the "mean" bowling score? Curtis N. Cooper and Robert E. Kennedy answer your question in

two articles: "Is the Mean Bowling Score Awful?" and "A Generating Function for the Distribution of the Scores of All Possible Bowling Games." It appears that the "mean" bowling score is lower than you may think, so perhaps your own score is not so bad.

Jim Propp analyzes a completely different type of game in "A New Take-Away Game." Suppose two children take turns stealing cookies from a cookie jar, each removing a single cookie every other day. Since some of the cookies may spoil before they are eaten, each cookie has an expiration date written on it in frosting. The goal of the game is to get the last edible cookie. You might want to play the game for a while before reading Propp's analysis. It turns out that there are many intriguing patterns involved, and much remains to be discovered before the game is completely analyzed.

A classical geometrical figure, first described by Archimedes in his *Book of Lemmas*, is the arbelos, the figure bounded by the arcs of three mutually tangent semicircles with their centers on the same straight line. [**4**, p. 304 ff.] This figure has many amazing properties, some of which were known to the Greeks, but many of which have been discovered recently. Leon Bankoff describes some of these properties in "The Marvelous Arbelos," but allows the reader the pleasure of discovering most of the proofs.

Richard Guy also challenges us with "The Strong Law of Small Numbers." There are many number patterns which seem to appear when we look at small values of n. But, as Guy notes, "there aren't enough small numbers to meet the many demands made of them." Thus many of the patterns—and we need to discover which ones—do not hold in general. If you like numerical puzzles, this article will provide many hours of stimulation (provided you don't look at the answers provided).

Finally, Lee C. F. Sallows teaches us about "Alphamagic Squares" in two articles. An alphamagic square is an ordinary magic square which remains magic when all of its entries are replaced by numbers representing the word length, in letters, of their conventional written names. For example, the integer 1 in a magic square is replaced by 3, the number of letters in "one." (Of course, an alphamagic square in English will not necessarily be an alphamagic square in French.) The skeptical reader may wonder if there are any alphamagic squares at all; but a quick check of the ancient Northumbrian *Li shu* square

$$\begin{array}{ccc} 5 & 22 & 18 \\ 28 & 15 & 2 \\ 12 & 8 & 25 \end{array}$$

shows that it is in fact alphamagic. Can you find others?

The Lighter Side of Mathematics will provide hours of challenge and "recreation" for those interested in exploring topics outside of the mainstream of academic mathematics. And faculty members and students alike may find that the "recreation" often raises very serious and interesting mathematical questions. But even if it doesn't, it is certainly fun.

REFERENCES

1. Ascher, Marcia, *Ethnomathematics: A Multicultural View of Mathematical Ideas* (Pacific Grove, CA.: Brooks/Cole Publishing Co., 1991.
2. Ball, W. W. Rouse and Coxeter, H. S. M., *Mathematical Recreations & Essays* (12th edition) (Toronto: University of Toronto Press, 1974).

3. Chace, Arnold Buffum, *The Rhind Mathematical Papyrus* (Reston, VA.: National Council of Teachers of Mathematics, 1967).
4. Heath, Thomas L., *The Works of Archimedes with the Method of Archimedes* (New York: Dover Publications, 1953).
5. Libbrecht, Ulrich, *Chinese Mathematics in the Thirteenth Century: The Shu-shu chiu-chang of Ch'in Chiu-shao* (Cambridge: MIT Press, 1973).
6. Vogel, Kurt (trans.), *Neun Bücher arithmetischer Technik* (Braunschweig: F. Vieweg and Sohn, 1968).
7. Zaslavsky, Claudia, *Africa Counts* (Boston: Prindle, Weber & Schmidt, 1973).

Mathematics Department
University of the District of Columbia
4200 Connecticut Ave. N.W.
Washington, DC 20008
vkatz@udcvax.bitnet

"A proof tells us where to concentrate our doubts."

"Universities hirc professors the way some men choose wives—they want the ones the others will admire."

Morris Kline, *Why Professor Can't Teach*, St. Martin's Press, 1977, p. 92.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee and Paul Zorn**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Mathematics Appreciation, T(13). *Chaos Under Control: The Art and Science of Complexity* David Peak, Michael Frame. WH Freeman, 1994, xiv + 408 pp, $24.95 (P). [ISBN 0-7167-2429-4] Readable, popular introduction to fractals and chaos for the "not quantitatively predisposed." Rounds up the usual suspects: iterated function systems, logistic map, Mandelbrot set, fractal basin boundaries, self-organizing phenomena. Problem sets, software, and syllabus for a course based on this book are available from the authors. SK

Mathematics Appreciation, L.** *Making Waves: A Guide to the Ideas Behind Outside In.* Silvio Levy. AK Peters, 1995, 49 pp, $39.95 (P), with video. [ISBN 1-56881-046-6] Geometry Center video uses computer graphics to illustrate an eversion of the sphere. Very accessible. Excellent use of graphics. Accompanying volume introduces background mathematics. Excellent addition to any library. TR

Precalculus, T(13: 1). *Precalculus: A Graphing Approach.* Dale Varberg, Thomas D. Varberg. Prentice Hall, 1995, xiv + 514 pp. [ISBN 0-13-010703-4] Uses graphing calculators to illustrate, illuminate traditional topics. Deemphasizes some manipulative skills in favor of visualization, modeling. Exercises in each section range from skills through applications to challenging problems suitable for projects. JCS

Finite Mathematics, T(13–14: 1).** *Finite Mathematics and Its Applications, Instructor's Edition, Fifth Edition.* Larry J. Goldstein, David I. Schneider, Martha J. Siegel. Prentice Hall, 1995, xvii + 802 pp. [ISBN 0-13-082778-9] Readable, student-friendly text covers linear programming, probability and statistics, game theory, graph theory, modeling, and more. Revised for more flexibility in course syllabus; probability and statistics chapters revised for readability. Many new exercises. Worked examples give ample detail for students who need review. Software to accompany text is available. (*Fourth Edition*, TR, April 1993.) LB

History, P. *Abraham Robinson: The Creation of Nonstandard Analysis: A Personal and Mathematical Odyssey.* Joseph Warren Dauben. Princeton Univ Pr, 1995, xix + 559 pp, $49.50. [ISBN 0-691-03745-0] Detailed biography combines explanations of Robinson's work in pure and applied mathematics with chronological story of his life. Includes information on his teachers, collaborators, and doctoral students. KES

History, P, L*. *The Infamous Boundary: Seven Decades of Controversy in Quantum Physics.* David Wick. Birkhäuser Boston, 1995, xii + 244 pp, $49.50. [ISBN 0-8176-3785-0] An overview of the history of quantum mechanics written for a general audience. Its main strength is the extended discussion of Bell's inequality and how it has been tested. DB

Foundations, T(14: 1), S, L. *The Keys to Advanced Mathematics: Recurrent Themes in Abstract Reasoning.* Daniel Solow. BookMasters Distrib Ctr (1444 US Route 42, RD #11, Mansfield, OH 44903), 1995, xx + 476 pp, $39.95 (P). [ISBN 0-9644519-0-5] Distinctive text for a "transition" course. First two chapters describe mathematical processes: unification, general-

ization, abstraction, development of axiomatic systems; and skills: writing proofs, identifying similarities and differences, converting visual images to symbolic form, understanding definitions. Chapters 3–6 apply these concepts to discrete mathematics, linear algebra, abstract algebra, and real analysis. KES

Combinatorics, P. *Surveys in Combinatorics, 1995.* Ed: Peter Rowlinson. London Math. Soc. Lect. Note Ser., V. 218. Cambridge Univ Pr, 1995, vii + 231 pp, $39.95 (P). [ISBN 0-521-49797-3] 9 survey articles from the 1995 British Combinatorial Conference.

Combinatorics, T(17: 1), P, L. *Extremal Combinatorial Problems and Their Applications.* Boris S. Stechkin, Valeriy I. Baranov. Math. & Its Applic., V. 335. Kluwer Academic, 1995, xiv + 205 pp, $110. [ISBN 0-7923-3631-3] Extremal problems in partitions, graphs, and geometry, with applications to computer memory allocation. Translated from Russian. TH

Number Theory, P. *Seminar on Fermat's Last Theorem.* Ed: V. Kumar Murty. Canadian Math. Soc. Conf. Proc., V. 17. AMS, 1995, x + 265 pp, $49 (P). [ISBN 0-8218-0313-1] Proceedings of a 1993–94 seminar at the Fields Institute in Toronto.

Number Theory, P. *Number Theory.* Ed: Karl Dilcher. Canadian Math. Soc. Conf. Proc., V. 15. AMS, 1995, xxiv + 431 pp, $89 (P). [ISBN 0-8218-0312-3] Proceedings of the Fourth Conference of the Canadian Number Theory Association (July 1994, Dalhousie University).

Linear Algebra, T(15–16: 1, 2). *Linear Algebra and Matrix Theory.* Jimmie Gilbert, Linda Gilbert. Academic Pr, 1995, x + 394 pp, $39.95. [ISBN 0-12-282970-0] For a second course (or a sophisticated first course). Emphasizes fundamental algebraic properties. Starts with $\mathbb{R}^n$ as a vector space; proceeds to abstract vector spaces, linear transformations, eigenvalues and eigenvectors, functions of vectors, inner product spaces, spectral decompositions, and the Jordan canonical form. Includes both computational and theoretical exercises. AO

Group Theory, P, L. *$SL(2)$ Representations of Finitely Presented Groups.* G.W. Brumfiel, H.M. Hilden. Contemp. Math., V. 187. AMS, 1995, vii + 196 pp, $49 (P). [ISBN 0-8218-0416-2] Classical theory with new results; makes connections to knot groups, invariant theory, three-dimensional manifolds. TH

Group Theory, P. *Representations of Groups.* Eds: Bruce N. Allison, Gerald H. Cliff. Canadian Math. Soc. Conf. Proc., V. 16. AMS, 1995, ix + 385 pp, $110 (P). [ISBN 0-8218-0311-5] Proceedings of the 1994 Annual Seminar of the Canadian Mathematical Society.

Group Theory, P. *Finite and Locally Finite Groups.* Eds: B. Hartley, *et al.* NATO ASI Ser. C, V. 471. Kluwer Academic, 1995, xii + 458 pp, $218. [ISBN 0-7923-3669-0] Proceedings of a 1994 NATO Advanced Study Institute held in Istanbul.

Algebra, P, L. *A Guide to Quantum Groups.* Vyjayanthi Chari, Andrew Pressley. Cambridge Univ Pr, 1994, xv + 651 pp, $34.95 (P); $64.95. [ISBN 0-521-55884-0; 0-521-43305-3] Comprehensive treatment of quantum groups and their applications. An ideal guide for both experts and newcomers. Emphasizes Lie algebras of quantum groups. Extensive bibliography. TH

Algebra, S(18), P, L. *Lectures on Lie Groups and Lie Algebras.* Roger Carter, Graeme Segal, Ian Macdonald. London Math. Soc. Stud. Texts, V. 32. Cambridge Univ Pr, 1995, viii + 190 pp, $22.95 (P); $39.95. [ISBN 0-521-49922-4; 0-521-49579-2] Based on short courses given at Lancaster University in 1993. First course, by Carter, covers Lie algebras and root systems. Outlines classification for simple Lie algebras over $\mathbb{C}$; discusses representations and simple groups of Lie type. Second course, by Segal, is on Lie groups. Includes compact groups, Peter–Weyl and Borel–Weil theorems, and representations. Third course, by Macdonald, is on linear algebraic groups. Starts from basics to work through Lie algebras, homogeneous spaces, maximal tori, and root structure. Bibliography, index. JS

Algebra, T(17: 1), L. *Groups and Representations.* J.L. Alperin, Rowen B. Bell. Grad. Texts in Math., V. 162. Springer-Verlag, 1995, x + 194 pp, $49.95. [ISBN 0-387-94526-1] First-year graduate course in group theory and representation theory of finite-dimensional algebras and finite groups. Emphasizes general linear groups. Includes Maschke and Wedderburn theorems, character theory. TH

Algebra, T(13: 1, 2). *College Algebra and Trigonometry.* Lewis Hirsch, Arthur Goodman. Prentice Hall, 1995, xix + 941 pp. [ISBN 0-13-031725-X] Traditional topics; highlights graphical, numerical, algebraic points of view. Graphing calculator exercises. TH

Algebra, P. *Noncommutative Algebraic Geometry and Representations of Quantized Algebras.* Alexander L. Rosenberg. Math. & Its Applic., V. 330. Kluwer Academic, 1995, xii + 315 pp, $155. [ISBN 0-7923-3575-9]

Algebra, P. *Near-Rings and Near-Fields.* Eds: Yuen Fong, *et al.* Math. & Its Applic., V. 336. Kluwer Academic, 1995, ix + 278 pp, $145. [ISBN 0-7923-3635-6] Proceedings of a 1993 conference in Fredericton, New Brunswick.

Algebra, T(15–16: 1, 2), S, L. *A Concrete Introduction to Higher Algebra, Second Edition.* Lindsay N. Childs. Undergrad. Texts in Math. Springer-Verlag, 1995, xv + 522 pp, $49.95. [ISBN 0-387-94484-2] New edition of a durable and well-regarded text. Admirers needn't feel concern—the essential elements and appeal of the *First Edition* have survived intact despite substantial revision (TR, October 1979; Extended Review, October 1983). Changes include more emphasis on elementary group theory, more treatment of computational number theory and primality testing, new coverage of quadratic reciprocity, use of the discrete Fourier transform. JS

Calculus, S(13). *The Derive Calculus Workbook.* Lisa Townsley Kulich, Barbara Victor. HarperCollins, 1994, xv + 239 pp, $17 (P). [ISBN 0-673-99455-4] 66 laboratory projects. Most on standard topics in single variable calculus; a few on multivariable topics. Labs are detailed and modest in scope, with much attention to *Derive* keystrokes. Includes 50-page guide to *Derive*. TR

Calculus, T(14). *Multivariable Calculus, Second Edition..* Leonard I. Holder, James DeFranza, Jay M. Pasachoff. Brooks/Cole, 1995, xvi + 605 pp, $62.25. [ISBN 0-534-24912-4] New features of this edition include calculator and CAS problems (identified by icons), interviews of calculus consumers in each chapter (including an economist, a physicist, and a medical researcher), illustrations with *Maple* and *Mathematica.* SK

Real Analysis, T(15). *An Introduction to Analysis.* William R. Wade. Prentice Hall, 1995, ix + 565 pp. [ISBN 0-13-093089-X] Clear, but rather dry, treatment of the theory of calculus, first in one variable and then in several. Includes chapters on Fourier series and on Stokes' theorem on manifolds. BH

Complex Analysis, P. *Lie Group Actions in Complex Analysis.* Dmitri N. Akhiezer. Aspects of Math., V. E27. Friedr Vieweg & Sohn, 1995, vii + 201 pp, $49. [ISBN 3-528-06420-X] Focuses on "the double role of Lie groups in complex analysis, namely as groups of biholomorphic selfmaps of certain complex analytic objects and as a special class of complex manifolds with an additional strong structure." BH

Differential Equations, P. *Modern Elementary Differential Equations, Second Edition.* Richard Bellman, Kenneth L. Cooke. Dover, 1995, xii + 228 pp, $8.95 (P). [ISBN 0-486-68643-4] Unabridged republication of the 1971 Addison-Wesley *Second Edition* (TR, May 1972).

Partial Differential Equations, P. *Mixed Problem for Partial Differential Equations with Quasihomogeneous Principal Part.* S.G. Gindikin, L.R. Volevich. Transl. of Math. Mono., V. 147. AMS, 1996, xi + 233 pp, $99. [ISBN 0-8218-4617-5]

Partial Differential Equations, P. *Fully Nonlinear Elliptic Equations.* Luis A. Caffarelli, Xavier Cabré. Colloquium Public., V. 43. AMS, 1995, vi + 104 pp, $39. [ISBN 0-8218-0437-5]

Partial Differential Equations, T(16–17), P. *A First Course in Partial Differential Equations with Complex Variables and Transform Methods.* H.F. Weinberger. Dover, 1995, ix + 446 pp, $14.95 (P). [ISBN 0-486-68640-X] An unabridged republication of the 1965 Blaisdell edition (TR, October 1973).

Dynamical Systems, T(16–17: 1, 2), P. *Elements of Applied Bifurcation Theory.* Yuri A. Kuznetsov. Appl. Math. Sci., V. 112. Springer-Verlag, 1995, xv + 515 pp, $59.95. [ISBN 0-387-94418-4] An introduction to nonlinear dynamical systems for applied mathematicians and scientists. Assumes only modest background in geometry, linear algebra, analysis, and differential equations. Includes numerical techniques for bifurcation analysis. AO

Numerical Analysis, P. *Infinite Element Methods.* Lung-an Ying. Friedr Vieweg & Sohn, 1995, 209 pp, DM 88. [ISBN 3-528-06610-5]

Numerical Analysis, P. *Solving Least Squares Problems.* Charles L. Lawson, Richard J. Hanson. Classics in Appl. Math., V. 15. SIAM, 1995, xii + 337 pp, $33 (P). [ISBN 0-89871-356-0] Unabridged, revised republication of the 1974 Prentice Hall edition (TR, June–July 1975). A new appendix summarizes major developments since first publication.

Numerical Analysis, T(16–17), P. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* Uri M. Ascher, Robert M.M. Mattheij, Robert D. Russell. Classics in Appl. Math., V. 13. SIAM, 1995, xxv + 595 pp, $39.50 (P). [ISBN 0-89871-354-4] Unabridged, corrected republication of the 1988 Prentice Hall edition.

Numerical Analysis, P*, L. *One Dimensional Spline Interpolation Algorithms.* Helmuth Späth. AK Peters, 1995, x + 404 pp,

$59.95. [ISBN 1-56881-016-4] Up-to-date, practical, introductory survey. Numerous examples and Fortran–77 subroutines make this book especially useful. AO

Numerical Analysis, T(18), P. *Iterative Methods for Linear and Nonlinear Equations.* C.T. Kelley. Frontiers in Appl. Math., V. 16. SIAM, 1995, xiii + 165 pp, $32.50 (P). [ISBN 0-89871-352-8] In-depth coverage of the conjugate gradient and generalized minimum residual techniques. Also discusses Newton–Krylov methods, new proofs and implementations of Broyden's method, and globalization of inexact Newton methods. AO

Operator Theory, P. *Theory of Commuting Nonselfadjoint Operators.* M.S. Livšic, *et al.* Math. & Its Applic., V. 332. Kluwer Academic, 1995, xvii + 313 pp, $174. [ISBN 0-7923-3588-0]

Functional Analysis, P. *Interaction Between Functional Analysis, Harmonic Analysis, and Probability.* Eds: Nigel Kalton, Elias Saab, Stephen Montgomery-Smith. Lect. Notes in Pure & Appl. Math., V. 175. Marcel Dekker, 1996, xii + 472 pp, $185. [ISBN 0-8247-9611-X] Proceedings of a 1994 conference at the University of Missouri.

Functional Analysis, P. *Function Spaces, The Second Conference.* Ed: Krzysztof Jarosz. Lect. Notes in Pure & Appl. Math., V. 172. Marcel Dekker, 1995, xi + 389 pp, $165 (P). Proceedings of a 1994 conference at Southern Illinois University at Edwardsville.

Functional Analysis, P. *Banach Space Complexes.* Călin-Grigore Ambrozie, Florian-Horia Vasilescu. Math. & Its Applic., V. 334. Kluwer Academic, 1995, 205 pp, $110. [ISBN 0-7923-3630-5]

Functional Analysis, T(17–18: 2). *Real Functions, Abstract Spaces and Orthogonal Series.* Miklós Mikolás. Académiai Kiadó, 1994, 494 pp, $69. [ISBN 963-05-6652-4] Designed to bridge gap between undergraduate real analysis and the modern study of functional analysis. Covers basic measure theory and integrability, and then goes into considerable depth in studying function spaces and orthogonal expansions. Translated and expanded from the Hungarian. No exercises. DB

Functional Analysis, P*. *Absolutely Summing Operators.* Joe Diestel, Hans Jarchow, Andrew Tonge. Stud. in Adv. Math., V. 43. Cambridge Univ Pr, 1995, xv + 474 pp, $59.95. [ISBN 0-521-43168-9] Aimed at those with solid backgrounds in real analysis, complex analysis, and functional analysis. Beautifully motivated exposition! Extensive bibliography. BH

Functional Analysis, P. *Generalized Topological Degree and Semilinear Equations.* Wolodymyr V. Petryshyn. Tracts in Math., V. 117. Cambridge Univ Pr, 1995, x + 240 pp, $59.95 (P). [ISBN 0-521-44474-8]

Functional Analysis, P. *Wavelets and Multiscale Signal Processing.* Albert Cohen, Robert D. Ryan. Appl. Math. & Math. Comput., V. 11. Chapman & Hall, 1995, vii + 238 pp. [ISBN 0-412-57590-6] A rigorous, clearly written introduction to wavelets and their use in digital signal processing. DB

Functional Analysis, T(17), P, L. *Applied Functional Analysis: Applications to Mathematical Physics.* Eberhard Zeidler. Appl. Math. Sci., V. 108. Springer-Verlag, 1995, xxix + 479 pp, $65. [ISBN 0-387-94442-7] First of two volumes. Rigorous introduction to Hilbert space theory starting from scratch. Development of theory richly interspersed with physics applications, including heat and wave equations, Schrödinger equation, quantum mechanics. Few problems. BH

Analysis, P. *Classification and Approximation of Periodic Functions.* Alexander I. Stepanets. Math. & Its Applic., V. 333. Kluwer Academic, 1995, x + 360 pp, $199. [ISBN 0-7923-3603-8]

Analysis, P. *Mathematical Analysis, Wavelets, and Signal Processing.* Eds: Mourad E.H. Ismail, *et al.* Contemp. Math., V. 190. AMS, 1995, x + 354 pp, $59 (P). [ISBN 0-8218-0384-0] Proceedings of a 1994 conference at Cairo University.

Analysis, P. *A Treatise on the Theory of Bessel Functions, Second Edition.* G.N. Watson. Cambridge Univ Pr, 1995, vii + 804 pp, $29.95 (P). [ISBN 0-521-48391-3] Republication of the 1944 *Second Edition.*

Analysis, P. *Variational Methods for Eigenvalue Problems: An Introduction to the Methods of Rayleigh, Ritz, Weinstein, and Aronszajn.* S.H. Gould. Dover, 1995, xii + 179 pp, $7.95 (P). [ISBN 0-486-68712-0] Unabridged republication of the 1957 University of Toronto Press edition.

Analysis, P. *Generalised Euler–Jacobi Inversion Formula and Asymptotics Beyond All Orders.* V. Kowalenko, *et al.* London Math. Soc. Lect. Note Ser., V. 214. Cambridge Univ Pr, 1995, xi + 129 pp, $29.95 (P). [ISBN 0-521-49798-1]

Algebraic Geometry, P. *Covariant Formal Group Theory and Some Applications.* I. Helsloot. CWI Tract, V. 111. Centrum voor

Wiskunde en Informatica, 1995, vii + 105 pp, Dfl. 35 (P). [ISBN 90-6196-455-5]

Geometry, P. *Invariant Methods in Discrete and Computational Geometry.* Ed: Neil L. White. Kluwer Academic, 1995, xiii + 328 pp, $170. [ISBN 0-7923-3548-1] Proceedings of the 1994 Curaçao Conference.

Algebraic Topology, T*(17–18: 1, 2), P, L. *Introduction to Differential and Algebraic Topology.* Yuri G. Borisovich, *et al.* Texts in Math. Sci., V. 9. Kluwer Academic, 1995, ix + 492 pp, $259. [ISBN 0-7923-3499-X] Captures reader's interest quickly by outlining problems that motivated development of topology, with links to other mathematical disciplines. Following basics of general topology, covers homotopy, manifolds, homology. Reviews pertinent literature for each chapter. Readable. Apart from cost, an excellent graduate text. LB

Optimization, P. *Stochastic Scheduling and Dynamic Programming.* G.M. Koole. CWI Tract, V. 113. Centrum voor Wiskunde en Informatica, 1995, 133 pp, Dfl. 35 (P). [ISBN 90-6196-457-1]

Optimal Control, P. *Lecture Notes in Control and Information Sciences–206: Linear Optimal Control of Bilinear Systems with Applications to Singular Perturbations and Weak Coupling.* Zijad Aganović, Zoran Gajić. Springer-Verlag, 1995, x + 133 pp, $45 (P). [ISBN 3-540-19976-4]

Probability, S(17), P, L. *Probability Theory: An Advanced Course.* Vivek S. Borkar. Universitext. Springer-Verlag, 1995, xiv + 138 pp, $29.95 (P). [ISBN 0-387-94558-X] Assumes background in applied probability and real analysis. Topics: probability measures, martingales, basic limit theorems, Markov chains, continuous-time processes. Designed for graduate students and researchers; most useful in planning research in stochastic processes. KB

Probability, T(17–18), P. *Asymptotic Methods in the Theory of Gaussian Processes and Fields.* Vladimir I. Piterbarg. Transl. of Math. Mono., V. 148. AMS, 1996, xii + 206 pp, $99. [ISBN 0-8218-0423-5] Studies asymptotic behavior of various functionals of Gaussian random variables and fields. Good approximation formulas and remainder estimates are obtained for many Gaussian and similar processes. No exercises. KB

Stochastic Processes, T(16–18: 3), P, L. *Applied Stochastic Processes: A Biostatistical and Population Oriented Approach.* Suddhendu Biswas. Wiley, 1995, xv + 427 pp, $39.95. [ISBN 0-470-22159-3] Theoretical approach to applied stochastic processes (random walks and Markov processes, non-Markov processes and renewal theory, martingales, theory of Geiger Muller counters and Palm probability). Applications treat epidemics, clinical drug trials, demography, population growth, survival and competing risk theory, genetics. No exercises. KB

Stochastic Processes, T(17–18), L. *Stochastic Differential Equations: An Introduction with Applications, Fourth Edition.* Bernt Øksendal. Springer-Verlag, 1995, xvi + 271 pp, $36 (P). [ISBN 0-387-60243-7] New material includes martingale representation theory, variational inequalities related to optimal stopping, stochastic control with terminal conditions. (*Second Edition*, TR, March 1990.) KB

Stochastic Processes, P. *Stochastic Partial Differential Equations.* Ed: Alison Etheridge. London Math. Soc. Lect. Notes, V. 216. Cambridge Univ Pr, 1995, x + 337 pp, $39.95 (P). [ISBN 0-521-48319-0] 19 papers from a March 1994 workshop at the University of Edinburgh.

Elementary Statistics, T(13–14: 1, 2), C. *An Introduction to Statistics for Managers.* Ann Spooner, Colin Lewis. Prentice Hall, 1995, xv + 398 pp, (P), with disk. [ISBN 0-13-030271-6] Basic introduction to statistics for business and management students. Covers basic probability, descriptive statistics, confidence intervals, hypothesis testing (means, variances, frequencies), simple linear regression and correlation. Applications include time series and forecasting, stock control, statistical process control. Closely tied to accompanying software (diskette included with text). LB

Statistics, T(13: 3, 4), C. *Business Statistics by Example, Fifth Edition.* Terry Sincich. Prentice Hall, 1996, xx + 1259 pp, [ISBN 0-02-410441-8]; *Computer Software Tutorials*, iii + 92 pp, (P), with disk. [ISBN 0-13-531609-X] Introductory text for economics and business majors with algebra background. Covers usual topics, plus time series, categorical data analysis, nonparametric statistics, decision analysis. Separate computer manual has easy-to-follow instructions for SAS, SPSS, MINITAB, ASP. KB

Statistics, P. *New Cambridge Statistical Tables, Second Edition.* D.V. Lindley, W.F. Scott. Cambridge Univ Pr, 1995, 96 pp, $12.95 (P). [ISBN 0-521-48485-5] From the Preface: "The only change from the *First Edition* is the inclusion of tables of Bayesian confidence intervals for the binomial and Poisson distributions and for the square of a multiple cor-

relation coefficient." (*First Edition*, TR, April 1985.)

Statistics, S(13), L. *How to Think about Statistics, Fifth Edition.* John L. Phillips, Jr. WH Freeman, 1996, xiv + 191 pp, $16.95 (P). [ISBN 0-7167-2822-2] Major change from previous edition (*Third Edition*, TR, December 1988) is emphasis on distinction between description and inference. New chapter treats transition from former to latter. Good preparation supplement for first statistics course. KB

Mathematical Computing, S(15–17), L*, P. *Solving Problems in Scientific Computing Using Maple and MATLAB, Second, Expanded Edition.* Walter Gander, Jiří Hřebíček. Springer-Verlag, 1995, xv + 315 pp, $39.95 (P). [ISBN 3-540-58746-2] Two new chapters on applications to electronic circuit design. Chapters on penetration phenomena and on metal compression have been expanded. (*First Edition*, TR, August–September 1994.) AO

Programming, P. *Schaum's Outline of Theory and Problems of Programming with Fortran 77.* William E. Mayo, Martin Cwiakala. Schaum's Outline Ser. McGraw-Hill, 1995, vi + 346 pp, $13.95 (P). [ISBN 0-07-041155-7]

Programming, P. *Schaum's Outline of Theory and Problems of Programming with Pascal, Second Edition.* Byron S. Gottfried. Schaum's Outline Ser. McGraw-Hill, 1994, viii + 428 pp, $13.95 (P). [ISBN 0-07-023924-X]

Programming, P. *Schaum's Outline of Theory and Problems of Programming with Fortran 90.* William E. Mayo, Martin Cwiakala. Schaum's Outline Ser. McGraw-Hill, 1996, viii + 440 pp, $14.95 (P). [ISBN 0-07-041156-5]

Programming, P. *C Programming FAQs: Frequently Asked Questions.* Steve Summit. Addison-Wesley, 1996, xli + 388 pp, $26.85 (P). [ISBN 0-201-84519-9]

Computer Systems, C, P. *Object-Oriented Design and Programming with C++.* Ronald Leach. Academic Pr, 1995, xiv + 463 pp, $34.95 (P), with disk. [ISBN 0-12-440215-1]

Computer Systems, P. *Programmer's Supplement for Release 6 of the X Window System, Version 11.* Ed: Adrian Nye. O'Reilly & Assoc, 1995, xxi + 430 pp, $29.95 (P). [ISBN 1-56592-089-9]

Computer Systems, P. *MH & xmh: Email for Users and Programmers, Third Edition.* Jerry Peek, Bill Wohler, Brent B. Welch. O'Reilly & Assoc, 1995, xlii + 738 pp, $34.95 (P). [ISBN 1-56592-093-7]

Computer Systems, P. *The Whole Internet for Windows 95: User's Guide & Catalog.* Ed Krol, Paula Ferguson. O'Reilly & Assoc, 1995, xix + 625 pp, $24.95 (P). [ISBN 1-56592-155-0]

Computer Systems, P. *Learning the bash Shell.* Cameron Newham, Bill Rosenblatt. O'Reilly & Assoc, 1995, xvi + 292 pp, $27.95 (P). [ISBN 1-56592-147-X]

Theory of Computation, P. *Loop Checking in Logic Programming.* R.N. Bol. CWI Tract, V. 112. Centrum voor Wiskunde en Informatica, 1995, 198 pp, Dfl. 40 (P). [ISBN 90-6196-456-3]

Applications (Biological Science), P. *Epidemic Models: Their Structure and Relation to Data.* Ed: Denis Mollison. Public. of Newton Inst. Cambridge Univ Pr, 1995, xvii + 424 pp, $49.95. [ISBN 0-521-47536-8] 23 papers from a 1993 NATO Advanced Research Workshop at the Newton Institute in Cambridge. In 5 parts: conceptual framework; spatial models; nonlinear time and space-time dynamics; heterogeneity in human diseases; data analysis—estimation and prediction.

Applications (Communication Theory), P. *Different Aspects of Coding Theory.* Ed: Robert Calderbank. Proc. of Symp. in Appl. Math., V. 50. AMS, 1995, viii + 239 pp, $49. [ISBN 0-8218-0379-4] Lecture notes for the 1995 AMS short course.

Applications (Engineering), P. *Schaum's Outline of Theory and Problems of Finite Element Analysis.* George R. Buchanan. Schaum's Outline Ser. McGraw-Hill, 1995, viii + 264 pp, $14.95 (P). [ISBN 0-07-008714-8]

Applications (Mechanics), T(17), L. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems.* Jerrold E. Marsden, Tudor S. Ratiu. Texts in Appl. Math., V. 17. Springer-Verlag, 1994, xv + 500 pp, $64.50. [ISBN 0-387-97275-7] A demanding introduction to Hamiltonian and Lagrangian mechanics. Includes basics of necessary differential geometry. Some exercises, many examples, references to current and recent literature. SK

Applications (Physics), P. *Dynamical Systems in Classical Mechanics.* Ed: V.V. Kozlov. Transl. Ser. 2, V. 168: Adv. in Math. Sci., 25. AMS, 1995, vii + 254 pp, $99. [ISBN 0-8218-0427-8] 10 papers from a seminar at Moscow Lomonosov University.

Reviewers

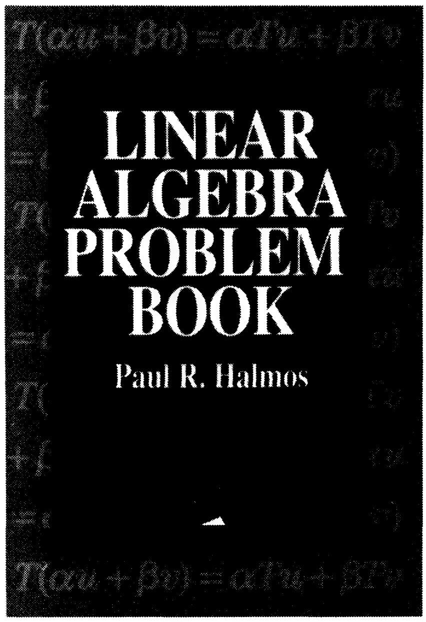
KB: Karla Ballman, Macalester; LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; TH: Tom Halverson, Macalester; BH: Bruce Hanson, St. Olaf; SK: Steve Kennedy, Carleton; AO: Arnold Ostebee, St. Olaf; TR: Tommy Ratliff, St. Olaf; JS: John Schue, Macalester; JCS: Janice Sklensky, St. Olaf; KES: Kay E. Smith, St. Olaf.

Linear Algebra Problem Book

Paul R. Halmos

Were it possible for the experience of apprenticeship to a master of mathematics to be packaged between the covers of a book, this would be it. No teacher of linear algebra should neglect to consult it. Highly recommended for all libraries.

– Choice Magazine



This is a book for mathematicians at all levels. Paul Halmos tells us, "Even if I know some answers, I don't think I understand a subject until I know the questions. The questions in mathematics are called problems—and although I learned some linear algebra a long time ago, until now I have made no serious effort to examine the problems that the solutions are based on. I wrote this book to organize those questions—problems—in my own mind."

This book is useful to anyone who needs linear algebra—and nowadays that means every user of mathematics. It can be used as the basis of either an official course or a program of private study.

If used as a course, the book can stand by itself, or if so desired, it can be stirred in with a standard linear algebra course as the seasoning that provides the interest, the challenge, the motivation that is needed by experienced scholars as much as by beginning students.

Contents

340 pp., Paperbound, 1995
ISBN 0-88385-322-1
List: $39.95
MAA Member: $28.95
Catalog Code: DOL-16/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-16/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

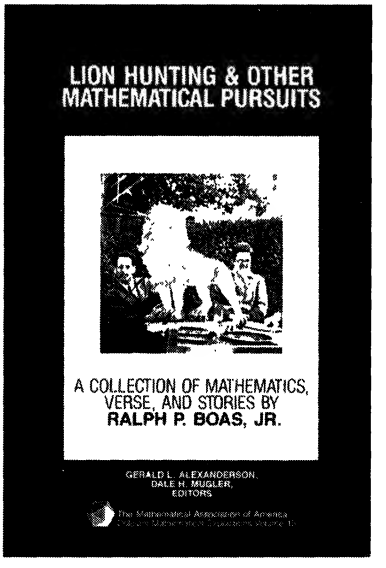
Signature ______________________

Lion Hunting

and Other Mathematical Pursuits

A Collection of Mathematics, Verse, and Stories
by Ralph P. Boas, Jr.

Gerald L. Alexanderson and
Dale H. Mugler, Editors



I highly recommend Lion Hunting and Other Mathematical Pursuits to high school mathematics clubs, mathematics teachers of all levels, and anyone interested in mathematics. Perhaps the most important features of this book is how it subtly makes the reader aware of the nature of mathematics.

– The Mathematics Teacher

As a young man at the Institute for Advanced Study in Princeton, Ralph Philip Boas, Jr., together with a group of other mathematicians, published a light-hearted article on the "mathematics of lion hunting" under a pseudonym (1938). This sparked a sequence of articles on the topic, several of which are drawn together in this book.

Lion Hunting includes an assortment of articles that show the many facets of this remarkable mathematician, editor, writer, and teacher. Along with a variety of his lighter mathematical papers, the collection includes Boas' verse and short stories, many of which are appearing for the first time. Anecdotes and recollections of his numerous experiences and of his work and meetings with many distinguished mathematicians and scientists of his day are also included as well as photographs taken by Boas of Hardy, Littlewood, Besicovitch, Weil, and others.

The mathematical articles in this collection cover a range of topics. They include articles on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, extremal problems for polynomials and more. A special section of this book is devoted to articles about the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we make mathematics intelligible?"

Boas's wit and playful humor are reflected in the verses included in this collection. The verses reflect the phases of his career as author, editor, teacher, department chair, and lover of literature. A section of the book describes the feud that Boas supposedly had with Bourbaki. Also included are many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995
ISBN 0-88385-323-X
List: $39.95
MAA Member: $28.95
Catalog Code: DOL-15/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-15/JR	______	______
______	______	______	______
		TOTAL	______

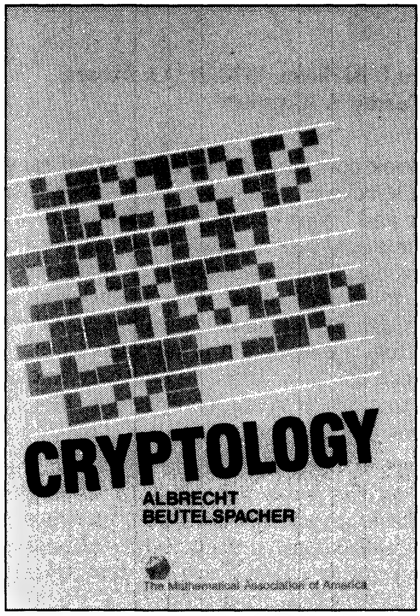
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________________ Expires ___/___

Signature ______________________

CRYPTOLOGY

Albrecht Beutelspacher



This fascinating little book is eminently readable, and it is a great deal of fun to peruse... the book is a real treat. We need more books like this, crafted by expert hands yet crafted so that the general reader can enjoy them.
—Bulletin of The Institute of Combinatorics and Its Applications

This excellent and entertaining book is suitable for a first course in cryptology for mathematical enthusiasts. An abundance of exercises and an excellent list of related references are included.
—The Mathematics Teacher

In spite of the light-hearted style in which the book is written throughout, it is a serious—-and successful—-attempt to explain the basis of coding and decoding messages...I can strongly recommend this book to anyone who wants a brief but comprehensive, eminently readable, and up-to-date introduction to this increasingly popular topic.
— The Mathematical Gazette

All of cryptology is covered in this work...Occupying a niche in the halls of the ivory tower of pure mathematics for nearly two millennia, number theory now forms a pillar of modern society. This book is the best explanation available today of how that pillar was constructed.
— Charles Aschbacher

A model to follow in order to make mathematics better known and understood. Accessible to a broad audience. Have fun reading this book, while you are getting a better understanding of cryptology.
— Bulletin of the Belgian Mathematics Society

How can messages be transmitted secretly? How can one guarantee that the message arrives safely in the right hands exactly as it was transmitted? Cryptology—the art and science of "secret writing"—provides ideal methods to solve these problems of data security.

The book is fun to read, and the author presents the material clearly and simply. Many exercises and references accompany each chapter.

176 pp., Paperbound, 1994
ISBN 0-88385-504-6
List: $34.00
MAA Member: $26.95
Catalog Code: CRYPT/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________________

Address ______________________________

City ______________________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CRYPT/JRw	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

Five Hundred Mathematical Challenges

Edward J. Barbeau, William O. Moser, and Murray S. Klamkin

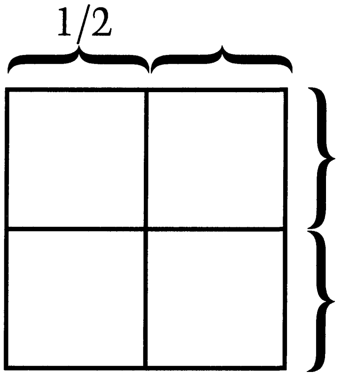
This book contains 500 problems that range over a wide spectrum of areas of high school mathematics and levels of difficulty. Some are simple mathematical puzzlers while others are serious problems at the Olympiad level. Students of all levels of interest and ability will be entertained and taught by the book. For many problems, more than one solution is supplied so that students can see how different approaches can be taken to a problem and compare the elegance and efficiency of different tools that might be applied.

Teachers at both the college and secondary levels will find the book useful, both for encouraging their students and for their own pleasure. Some of the problems can be used to provide a little spice in the regular curriculum by demonstrating the power of very basic techniques.

These problems were first published as a series of problem booklets almost twenty years ago, at a time when there were few resources of this type available for the English reader. They have stood the test of time and the demand for them has been steady. Their publication in book form is long overdue.

This collection provides a solid base for students who wish to enter competitions at the Olympiad level. They can begin with easy problems and progress to more demanding ones. A special mathematical tool chest summarizes the results and techniques needed by competition-level students.

Problem 14.
Show that if 5 points are all in, or on, a square of side 1, then some pair of them will be no further than $\frac{\sqrt{2}}{2}$ apart.



Solution.
Divide the square into four squares of side 1/2. By the Pigeonhole Principle, one of these four squares contains at least two of the points, whose distance apart must be no greater than the diagonal of the square of side 1/2, namely $\frac{\sqrt{2}}{2}$.

300 pp., Paperbound, 1995
ISBN 0-88385-519-4

List: $29.50
MAA Member: $24.95
Catalog Code: CHMP/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ___ ___ ___ ___ ___ ___

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CHMP/JR	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires __/__

Signature ______________________

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 103, Number 4 April 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

Daniel ULLMAN, MONTHLY problems
Department of Mathematics
The George Washington University
2201 G. Street, NW, Room 428B
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Similarities Between Fourier and Power Series

Richard Askey and Deborah Tepper Haimo

In Memory of Ralph P. Boas, Jr.

1. INTRODUCTION. In a paper titled "An Unorthodox Test" in the January 1992 issue of the MONTHLY, Abe Shenitzer of York University poses 16 questions that he feels are intellectually vital in the teaching of mathematics. The second of these asks:

> What are some basic differences between Taylor series and Fourier series?

In his response, in which he considers functions restricted to the reals, he points out that the infinite differentiability of a function does not itself assure that its power series will converge to that function, whereas mere periodicity and a little smoothness are enough to have the Fourier series converge uniformly to the function. Further, the terms of the Fourier series describe simple harmonic motion so that the function may be considered as a linear combination of harmonic motions, whereas the terms of a power series have no such physical interpretation.

While calling attention to such differences is instructive, an interesting and deeper question might also be posed.

> What are some basic similarities between power series and Fourier series beyond the fact that they are both infinite series?

One obvious answer to this question is that if the power series is

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad -r < x < r,$$

and if it is continued to the complex plane, then

$$f(s e^{i\theta}) = \sum_{n=0}^{\infty} a_n s^n e^{in\theta}, \qquad 0 \leq s < r,$$

is a Fourier series in θ. Like Shenitzer, however, we restrict ourselves to functions of a real variable.

In this context, we find that there is another connection that demonstrates more dramatic similarities, and at the same time, helps to explain some of the differences Shenitzer mentions. This is the existence of an expansion of a function f in a series of ultraspherical polynomials, $C_n^\lambda(x)$, $\lambda > 0$, that contains Fourier series of even functions and power series as special limiting cases. We establish this result by demonstrating that since the $C_n^\lambda(x)$ are orthogonal, the coefficients in the expansion can be represented by integrals, and if $\cos\theta$ is substituted for x, and we

let $\lambda \to 0$, the resulting expansion becomes a Fourier series. On the other hand, if f is infinitely differentiable, n integrations by parts can be applied to obtain new integral representations of the coefficients, so that for analytic functions, by letting $\lambda \to \infty$, we obtain the classical formula for the coefficients as nth derivatives in the power series expansion of a function.

2. FOURIER SERIES AND POWER SERIES. Power series of a real variable have the form

$$f(x) = \sum_{n=0}^{\infty} c_n x^n, \tag{2.1}$$

where the coefficients are given by

$$c_n = f^{(n)}(0)/n!. \tag{2.2}$$

Since we will consider only even Fourier series, we have

$$g(\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos n\theta \tag{2.3}$$

with

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\cos n\theta \, d\theta, \qquad n = 0, 1, \dots . \tag{2.4}$$

We note that the coefficients in a power series are given by the local formula (2.2) involving derivatives, while those in a Fourier series are represented by the global integral formula (2.4). The two are related to each other as we have noted and will demonstrate.

3. ULTRASPHERICAL POLYNOMIALS. In work on planetary motion, Legendre and Laplace introduced a set of orthogonal polynomials, $P_n(x)$, now called *Legendre polynomials* and generated by

$$(1 - 2xr + r^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x) r^n. \tag{3.1}$$

These were later extended to the so-called *Gegenbauer* or *ultraspherical polynomials*, $C_n^\lambda(x)$, given by

$$(1 - 2xr + r^2)^{-\lambda} = \sum_{n=0}^{\infty} C_n^\lambda(x) r^n. \tag{3.2}$$

An explicit formula that clearly indicates that n is the degree of $C_n^\lambda(x)$ is

$$C_n^\lambda(x) = \sum_{k=0}^{[n/2]} \frac{(-1)^k (\lambda)_{n-k}}{k!(n-2k)!} (2x)^{n-2k}, \tag{3.3}$$

where

$$(\alpha)_n = \Gamma(n+\alpha)/\Gamma(\alpha) \tag{3.4}$$

is the shifted factorial $\alpha(\alpha+1)\cdots(\alpha+n-1)$. This and other properties of the ultraspherical polynomials can be found in Chapter IV of **[6]** and Chapter 10 of **[1]**.

We may readily obtain a second representation of $C_n^\lambda(x)$ by setting $x = \cos\theta$ in (3.2), factoring

$$1 - 2r\cos\theta + r^2 = (1 - re^{i\theta})(1 - re^{-i\theta}),$$

expanding each factor on the right by the binomial theorem, and using the Cauchy product of two series. We find that

$$\begin{aligned} C_n^\lambda(\cos\theta) &= \sum_{k=0}^{n} \frac{(\lambda)_{n-k}(\lambda)_k}{(n-k)!k!} e^{i(n-2k)\theta} \\ &= \sum_{k=0}^{n} \frac{(\lambda)_{n-k}(\lambda)_k}{(n-k)!k!} \cos(n-2k)\theta, \end{aligned} \tag{3.5}$$

the last holding because $C_n^\lambda(\cos\theta)$ is real when λ is. A connection with Fourier series results from the fact that

$$\lim_{\lambda\to 0} \frac{C_n^\lambda(\cos\theta)}{\lambda} = \frac{2}{n}\cos n\theta, \qquad n = 1, 2, 3, \dots, \tag{3.6}$$

which follows from (3.5) since all except the end terms have two factors of λ. From this limit relation, we immediately arrive at two other forms that are more useful for us since they have the advantage of holding for $n = 0$ as well as for positive n, while (3.6) holds only for positive n. One is

$$\lim_{\lambda\to 0} \frac{C_n^\lambda(\cos\theta)}{C_n^\lambda(1)} = \cos n\theta, \qquad n = 0, 1, \dots, \tag{3.7}$$

and the other,

$$\lim_{\lambda\to 0} \frac{n+\lambda}{\lambda} C_n^\lambda(\cos\theta) = \begin{cases} 1 & n = 0 \\ 2\cos n\theta & n = 1, 2, \dots \end{cases}. \tag{3.8}$$

A different limiting case comes from (3.5). Since from (3.2) we have

$$\begin{aligned} \sum_{n=0}^{\infty} C_n^\lambda(1) r^n &= (1-r)^{-2\lambda} \\ &= \sum_{n=0}^{\infty} \frac{(2\lambda)_n}{n!} r^n, \end{aligned}$$

it follows that

$$C_n^\lambda(1) = \frac{(2\lambda)_n}{n!}. \tag{3.9}$$

Now from (3.5) and (3.9), we have

$$\begin{aligned} \lim_{\lambda\to\infty} \frac{C_n^\lambda(\cos\theta)}{C_n^\lambda(1)} &= \lim_{\lambda\to\infty} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \frac{(\lambda)_k(\lambda)_{n-k}}{(2\lambda)_n} e^{i(n-2k)\theta} \\ &= \frac{e^{in\theta}}{2^n} \sum_{k=0}^{n} \binom{n}{k} e^{-2ik\theta} \\ &= \frac{e^{in\theta}(1+e^{-2i\theta})^n}{2^n} \\ &= \cos^n\theta. \end{aligned} \tag{3.10}$$

Hence

$$\lim_{\lambda\to\infty} \frac{C_n^\lambda(x)}{C_n^\lambda(1)} = x^n. \tag{3.11}$$

We have now established that the building blocks of Fourier and of power series are limiting cases of ultraspherical polynomials. We determine the desired expansions in terms of these polynomials, well known to be orthogonal. Their orthogonality relation is given by

$$\int_{-1}^{1} C_n^\lambda(x) C_m^\lambda(x)(1-x^2)^{\lambda-1/2}\, dx = \frac{\lambda}{(n+\lambda)} \frac{(2\lambda)_n}{n!} A_\lambda\, \delta(m,n) \quad (3.12)$$

where

$$A_\lambda = \int_{-1}^{1} (1-x^2)^{\lambda-1/2}\, dx$$

$$= 2^{2\lambda} \frac{\Gamma(\lambda+1/2)^2}{\Gamma(2\lambda+1)} \quad (3.13)$$

and $\delta(m,n)$ vanishes when $m \neq n$ and is 1 when $m = n$.

We will need also the familiar *Rodrigues formula*, which for the ultraspherical polynomials is given by

$$(1-x^2)^{\lambda-1/2} C_n^\lambda(x) = \frac{(-1)^n (2\lambda)_n}{2^n(\lambda+1/2)_n n!} \frac{d^n}{dx^n}(1-x^2)^{n+\lambda-1/2}. \quad (3.14)$$

Further, an extension of the familiar trigonometric addition formula, established in [3] by Gegenbauer and given in integral form as

$$C_n^\lambda(\cos\theta) C_n^\lambda(\cos\phi)$$

$$= \frac{C_n^\lambda(1)}{A_{\lambda-1/2}} \int_0^\pi C_n^\lambda(\cos\theta\cos\phi + \sin\theta\sin\phi\cos\psi)(\sin\psi)^{2\lambda-1}\, d\psi, \quad (3.15)$$

reduces to

$$\cos n\theta \cos n\phi = [\cos n(\theta+\phi) + \cos n(\theta-\phi)]/2 \quad (3.16)$$

when $\lambda \to 0$, and becomes

$$(\cos\theta)^n (\cos\phi)^n = (\cos\theta\cos\phi)^n \quad (3.17)$$

when $\lambda \to \infty$.

4. A CONNECTION BETWEEN FOURIER AND POWER SERIES. We now turn to the orthogonal expansion in terms of ultraspherical polynomials that will provide the connection we seek between Fourier and power series. It is given by

$$f(x) = \sum_{n=0}^{\infty} a_n^{(\lambda)} \frac{(n+\lambda)}{\lambda} C_n^\lambda(x), \quad (4.1)$$

with $a_n^{(\lambda)}$ determined by

$$a_n^{(\lambda)} = \frac{1}{A_\lambda} \int_{-1}^{1} f(t) \frac{C_n^\lambda(t)}{C_n^\lambda(1)} (1-t^2)^{\lambda-1/2}\, dt \quad (4.2)$$

and A_λ defined in (3.13). Now, if we substitute $\cos\theta$ for x in (4.1) and $\cos\phi$ for t in (4.2), let $\lambda \to 0$, and set $a_n = \lim_{\lambda\to 0} a_n^{(\lambda)}$, we have, using (3.8) and (3.7),

respectively,

$$f(\cos\theta) = a_0 + 2\sum_{n=1}^{\infty} a_n \cos n\theta, \tag{4.3}$$

where

$$a_n = \frac{1}{\pi}\int_0^{\pi} f(\cos\phi)\cos n\phi\, d\phi. \tag{4.4}$$

The series (4.3) is a result of interchanging the limit as $\lambda \to 0$ with the summation in (4.1), and the integral (4.4) is a consequence of permuting that limit with the integral in (4.2), both justifiable.

We now have shown that the Fourier series of even functions are contained in the ultraspherical expansion (4.1) when $\lambda \to 0$.

To see how power series arise, we first change the integral representation (4.2) into one that involves the nth derivative of $f(x)$. Assuming that f is infinitely differentiable and applying the Rodrigues formula (3.14) to (4.2), we have, after n integrations by parts,

$$a_n = \frac{1}{A_\lambda 2^n(\lambda+1)_n}\int_{-1}^{1} f^{(n)}(x)(1-x^2)^{n+\lambda-1/2}\,dx. \tag{4.5}$$

It follows that (4.1) becomes

$$f(x) = \sum_{n=0}^{\infty} \frac{n+\lambda}{\lambda}\frac{(2\lambda)_n}{n!}\frac{1}{2^n(\lambda+1)_n}\frac{\int_{-1}^{1} f^{(n)}(t)(1-t^2)^{n+\lambda-1/2}\,dt}{\int_{-1}^{1}(1-t^2)^{n+\lambda-1/2}\,dt}\frac{C_n^\lambda(x)}{C_n^\lambda(1)}. \tag{4.6}$$

When $\lambda \to \infty$, and again the limit can be moved within the summation, (4.6) becomes

$$f(x) = \sum_{n=0}^{\infty}\frac{f^{(n)}(0)}{n!}x^n \tag{4.7}$$

since

$$\frac{C_n^\lambda(x)}{C_n^\lambda(1)} \to x^n \tag{4.8}$$

by (3.11),

$$\lim_{\lambda\to\infty}\frac{n+\lambda}{\lambda}\frac{(2\lambda)_n}{2^n(\lambda+1)_n} = 1 \tag{4.9}$$

and the measure

$$\frac{(1-x^2)^{n+\lambda-1/2}}{\int_{-1}^{1}(1-t^2)^{n+\lambda-1/2}\,dt}dx \tag{4.10}$$

is an approximation to the delta function. The latter results from the fact that it has mass one on $[-1,1]$, and its distribution becomes singular when $\lambda \to \infty$, with all of its mass concentrated at $x = 0$.

By showing that the power series also comes from ultraspherical expansions, in this case when $\lambda \to \infty$, we have established our result using the set of ultraspherical polynomials to bridge the gap between Fourier and power series.

5. SMOOTHNESS CONDITIONS AND CONVERGENCE.

5. SMOOTHNESS CONDITIONS AND CONVERGENCE. Shenitzer remarked that Fourier series converge when the function represented has a little smoothness, but infinite differentiability is not sufficient to ensure the convergence of a power series to the function expanded. Let us examine the convergence of ultraspherical series to get a clearer understanding of the reason for this. Without including full details, we outline an idea introduced by Lebesgue for finding sufficient conditions for the uniform convergence of the ultraspherical series.

We begin with the partial sums

$$S_n(f;x) = \sum_{k=0}^{n} a_k \frac{(k+\lambda)}{\lambda} C_k^\lambda(x)$$
$$= \frac{1}{A_\lambda}\int_{-1}^{1} f(y) \sum_{k=0}^{n} \frac{(k+\lambda)}{\lambda} \frac{k!}{(2\lambda)_k} C_k^\lambda(x) C_k^\lambda(y)(1-y^2)^{\lambda-1/2}\, dy.$$

Since, for an arbitrary polynomial $p_k(x)$ of degree k,

$$S_n(p_k;x) = p_k(x), \qquad n \geq k,$$

we note that

$$|f(x) - S_n(f;x)| = |f(x) - p_k(x) - S_n(f-p_k;x)|$$
$$\leq |f(x) - p_k(x)| + |S_n(f-p_k;x)|.$$

We have, however, that

$$|S_n(f-p_k;x)| \leq \max_{-1\leq x\leq 1} |f(x) - p_k(x)| \max_{-1\leq x\leq 1} \rho_n^\lambda(x), \tag{5.1}$$

where the $\rho_n^\lambda(x)$ are defined by

$$\rho_n^\lambda(x) = \frac{1}{A_\lambda}\int_{-1}^{1} \left| \sum_{k=0}^{n} \frac{(k+\lambda)}{\lambda} \frac{k!}{(2\lambda)_k} C_k^\lambda(x) C_k^\lambda(y) \right| (1-y^2)^{\lambda-1/2}\, dy. \tag{5.2}$$

To establish our result, we first show that

$$\rho_n^\lambda(x) \leq \rho_n^\lambda(1). \tag{5.3}$$

To this end, if we let $x = \cos\theta$, $y = \cos\phi$, and $z = \cos\theta\cos\phi + \sin\theta\sin\phi\cos\psi$, we can rewrite (3.15) in the form

$$C_n^\lambda(x) C_n^\lambda(y) = C_n^\lambda(1)\int_{-1}^{1} K_\lambda(x,y,z) C_n^\lambda(z)(1-z^2)^{\lambda-1/2}\, dz, \tag{5.4}$$

where

$$K_\lambda(x,y,z) = \frac{1}{A_{\lambda-1/2}}(1-x^2-y^2-z^2+2xyz)^{\lambda-1}$$
$$\times\left[(1-x^2)(1-y^2)(1-z^2)\right]^{1/2-\lambda}, \tag{5.5}$$

with $K_\lambda(x, y, z) = 0$ when $z < xy - (1-x^2)^{1/2}(1-y^2)^{1/2}$ or $z > xy + (1-x^2)^{1/2}(1-y^2)^{1/2}$.

We note from (5.4) that the kernel, $K_\lambda(x,y,z)$, may be represented by the series

$$K_\lambda(x,y,z) = \sum_{k=0}^{\infty} \frac{(k+\lambda)}{\lambda} C_k^\lambda(x) C_k^\lambda(y) C_k^\lambda(z)/C_k^\lambda(1). \tag{5.6}$$

It is non-negative, as is clear from (5.5), and

$$\int_{-1}^{1} K_\lambda(x, y, z)(1 - z^2)^{\lambda - 1/2}\, dz = 1, \tag{5.7}$$

as follows readily from (5.4) when $n = 0$.

Now, using (5.2), we establish (5.3). This follows from the fact that we have

$$\begin{aligned}
\rho_n^\lambda(x) &= \frac{1}{A_\lambda}\int_{-1}^{1}\left|\sum_{k=0}^{n}\frac{(k+\lambda)}{\lambda}\frac{k!}{(2\lambda)_k}C_k^\lambda(x)C_k^\lambda(y)\right|(1-y^2)^{\lambda-1/2}\,dy\\
&= \frac{1}{A_\lambda}\int_{-1}^{1}\left|\sum_{k=0}^{n}\frac{(k+\lambda)}{\lambda}\int_{-1}^{1}K_\lambda(x,y,z)C_k^\lambda(z)(1-z^2)^{\lambda-1/2}\,dz\right|(1-y^2)^{\lambda-1/2}\,dy\\
&\le \frac{1}{A_\lambda}\int_{-1}^{1}\left|\sum_{k=0}^{n}\frac{(k+\lambda)}{\lambda}C_k^\lambda(z)(1-z^2)^{\lambda-1/2}\right|\int_{-1}^{1}K_\lambda(x,y,z)(1-y^2)^{\lambda-1/2}\,dy\,dz\\
&= \frac{1}{A_\lambda}\int_{-1}^{1}\left|\sum_{k=0}^{n}\frac{(k+\lambda)}{\lambda}C_k^\lambda(z)\right|(1-z^2)^{\lambda-1/2}\,dz\\
&= \rho_n^\lambda(1).
\end{aligned}$$

Let us now use the estimate of the size of $\rho_n^\lambda(1)$ determined by H. Rau in [**5**]. He showed that, for each $\lambda > 0$, there is a positive constant B_λ such that

$$\rho_n^\lambda(1) = B_\lambda n^\lambda + o(n^\lambda). \tag{5.8}$$

It is a classical result that when $\lambda = 0$, the right-hand side of (5.8) becomes a constant times $\log n$ for the *Fourier series*.

Now, appealing to the inequality (5.1) as well as that immediately preceding it, and taking note of the bound for $\rho_n^\lambda(x)$ and Rau's result (5.8), we have, for $\lambda > 0$,

$$\begin{aligned}
|f(x) - S_n(f;x)| &\le |f(x) - p_k(x)| + \max_{-1\le t\le 1}|f(t) - p_k(t)| \max_{-1\le x\le 1}\rho_n^\lambda(t)\\
&\le A n^\lambda \max_{-1\le t\le 1}|f(t) - p_k(t)|,
\end{aligned}\tag{5.9}$$

for all $k \ge n$. The right-hand side of (5.9) can be estimated from classical results on polynomials and trigonometric polynomials. This can most readily be accomplished by taking $x = \cos\theta$ and asking what smoothness conditions on a function on the unit circle will result in

$$|f(\cos\theta) - p_k(\cos\theta)| = o(n^{-\lambda}). \tag{5.10}$$

If $g(\theta) = f(\cos\theta)$ has $[\lambda] = j$ continuous derivatives, and if the jth derivative of $g(\theta)$ satisfies

$$\sup_{0\le\theta\le\pi}|g^{(j)}(\theta+\phi) - g^{(j)}(\theta)| = o(|\phi|^\alpha), \qquad 0 \le \alpha \le 1, \tag{5.11}$$

then

$$|f(\cos\theta) - p_k(\cos\theta)| = o(n^{-j-\alpha}), \qquad 0 \le \theta \le \pi. \tag{5.12}$$

Now take $\alpha = \lambda - [\lambda]$. When $0 < \alpha < 1$, if (5.11) is satisfied, then $f(\cos\theta) = g(\theta)$ has j continuous derivatives and (5.10) holds. For $\alpha = 0$, a slightly weaker condition is satisfied. See Zygmund [**7**, Chapter 3, section 13].

The results for uniform convergence are best possible. Thus, the degree of smoothness, which implies uniform convergence, increases as λ does.

6. THE JACOBI POLYNOMIALS. All the above results may be extended by considering a more general set of orthogonal polynomials, the Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, defined by

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{n+\alpha}(1+x)^{n+\beta}\right], \quad (6.1)$$

which include the $C_n^{\lambda}(x)$ as the special case where $\alpha = \beta = \lambda - 1/2 > -1$. These polynomials satisfy slightly more complicated versions of all the formulas developed for the ultraspherical polynomials. For these polynomials, we have, for example, that

$$\frac{P_n^{(-1/2,-1/2)}(\cos\theta)}{P_n^{(-1/2,-1/2)}(1)} = \cos n\theta, \quad (6.2)$$

so that the even Fourier series occur directly and not as limiting cases. Further, we can also find sufficient conditions for the uniform convergence of the Jacobi series, establishing results analogous to those we have for the ultraspherical case. For this, we need an extension of the Gegenbauer formula for Jacobi polynomials, found by G. Gasper in **[2]**.

The arguments we developed above were largely known, the basic facts about properties of orthogonal polynomials having been discovered well before the 20th century. It is not uncommon in mathematics to have important results overlooked or forgotten over the years. The formulation of Shenitzer's question suggests that the close connection between Fourier and power series, provided by ultraspherical and Jacobi series, is not generally known now. Some classical analysts, on the other hand, including for example Szegő, **[6]**, were well aware of this useful relationship, and it has served as a source of interesting problems throughout the decades.

REFERENCES

1. Erdélyi, A. et al., *Higher Transcendental Functions*, vol. 2, McGraw Hill, N.Y., 1953.
2. Gasper, G., Positivity and the convolution structure for Jacobi series, *Annals of Math.* (2) 93 (1971), 112–118.
3. Gegenbauer, L., Über einige bestimmte Integrale, *Sitzungsberichte Kaiserlichen Akad. der Wiss. math.-natur. Klasse Wien*, 70 (2a) (1874), 433–443.
4. Gronwall, T. H., On the degree of convergence of Laplace series, *Trans. Amer. Math. Soc.* 15 (1914), 1–30.
5. Rau, H., Über die Lebesgueschen Konstanten der Reihenentwicklungen nach Jacobischen Polynomen, *J. für reine und angew. Math.* 161 (1929), 237–254.
6. Szegő, G., *Orthogonal Polynomials*, Amer. Math. Soc. Colloquium Publ. XXIII, fourth ed. (1975), Chapter IV.
7. Zygmund, A., *Trigonometric Series*, 2nd ed., Vol. I, Cambridge Univ. Press, N.Y., 1959.

Department of Mathematics
University of Wisconsin–Madison
Madison, WI 53706
askey@math.wisc.edu

Department of Mathematics, 0112
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
dhaimo@euclid.ucsd.edu

Disks on a Chessboard

Richard H. Warren

The puzzle in this note is similar to the one that was analyzed in [1]. After characterizing the configurations in our puzzle as representing cyclic permutations, we apply this result to the traveling salesman problem.

We begin with an infinite "chessboard" covering the first quadrant. The cells of the board are labelled by integer coordinates (i, j) with $i, j \geq 0$. Initially, a single "disk" is located in cell $(0, 0)$. The first step consists of replacing this disk by two disks, located in cells $(1, 0)$ and $(0, 1)$, respectively. In general, if there are k disks on the board, a step will consist of removing some disk, say in cell (i, j), and placing two disks on the board, one disk located in cell (k, j) and the other disk located in cell (i, k).

Lemma 1. *If there are k disks on the board, then none of the cells in row k or column k contains a disk.*

Lemma 2. *A cell contains at most one disk.*

After k steps the board will have $k + 1$ disks on it. We call such configurations of cells with disks *attainable configurations*. In Figure 1 we show the ten possible attainable configurations with at most four disks.

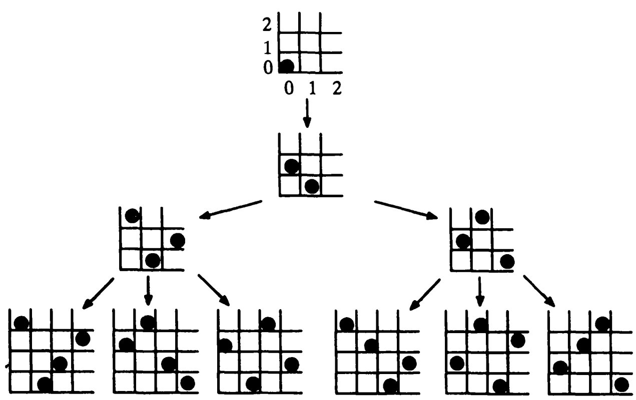


Figure 1. Attainable configurations with at most four disks.

It is easy to see that an attainable configuration with k disks has exactly one disk in each of the first k rows of the board and in each of the first k columns of the board. This means that an attainable configuration represents a permutation, in the same way that a permutation matrix represents a permutation. Recall that a cyclic permutation of k objects has exactly one cycle. For example, the cyclic

permutation of four objects represented by $0 \to 1 \to 3 \to 2 \to 0$ has one cycle denoted by (0132).

Theorem. *A configuration of cells is an attainable configuration if and only if the configuration represents a cyclic permutation.*

Proof: Let k cells be given where $k \geq 2$. Suppose that the k cells are an attainable configuration. If $k = 2$, then the attainable configuration represents $0 \to 1 \to 0$. Assume that if $k = n$, then the cells in an attainable configuration represent a cyclic permutation $0 \to a_1 \to \cdots \to a_{n-1} \to 0$. For $k = n + 1$, let the attainable configuration be formed by replacing cell (i, j) with cells (n, j) and (i, n). This means that the cyclic permutation $0 \to a_1 \to \cdots \to i \to n \to j \to \cdots \to a_{n-1} \to 0$ represents the configuration.

Conversely, suppose that a configuration of k cells represents a cyclic permutation of $\{0, 1, \dots, k - 1\}$. If $k = 2$, since the only cyclic permutation is $0 \to 1 \to 0$, the cyclic permutation is represented by the attainable configuration that has cells $(1, 0)$ and $(0, 1)$. For $k = n$, the induction assumption is that if the given cells represent a cyclic permutation, then the configuration of the cells is an attainable configuration. For $k = n + 1$, let the given cells represent a cyclic permutation that contains $\cdots \to i \to n \to j \to \cdots$. We delete n from the cyclic permutation. The result is a cyclic permutation of $\{0, 1, \dots, n - 1\}$ corresponding to an attainable configuration that includes cell (i, j). We replace cell (i, j) with cells (n, j) and (i, n). This forms an attainable configuration that is the original configuration. ■

Let $\{0, 1, 2, \dots, n\}$ designate a set of cities. Let d_{ij} be the distance from city i to city j. The traveling salesman problem (TSP) asks if the following question can be decided in a polynomial number of steps, where the variable in the polynomial is n. Given an integer K, is there a cyclic permutation σ of $\{0, 1, 2, \dots, n\}$ such that

$$d_{0\sigma(0)} + d_{1\sigma(1)} + d_{2\sigma(2)} + \cdots + d_{n\sigma(n)} \leq K?$$

The question is equivalent to asking: Is there an algorithm with a polynomial number of steps, where the variable in the polynomial is n, that will find a cyclic permutation σ of $\{0, 1, 2, \dots, n\}$ such that the sum

$$d_{0\sigma(0)} + d_{1\sigma(1)} + d_{2\sigma(2)} + \cdots + d_{n\sigma(n)}$$

is a minimum over the sums for all cyclic permutations of $\{0, 1, 2, \dots, n\}$?

Since the answer to the question is not known, a large percent of the work on the TSP is about "fast" methods to find a cyclic permutation of the cities such that the sum of the distances is "close" to the minimum. The TSP has numerous applications and has received wide publicity [2]–[9].

The induction technique used to analyze the puzzle often applies to the TSP when proving properties about the salesman's tour through the cities (i.e., about cyclic permutations). The technique is to assume the property is true for a set S_k of cyclic permutations of k cities. A set S_{k+1} of cyclic permutations of $k + 1$ cities is formed by inserting a new city between cities i and j in each member of S_k. Each cyclic permutation in S_k gives rise to k cyclic permutations in S_{k+1}. Then it is necessary to prove the property for the members of S_{k+1}.

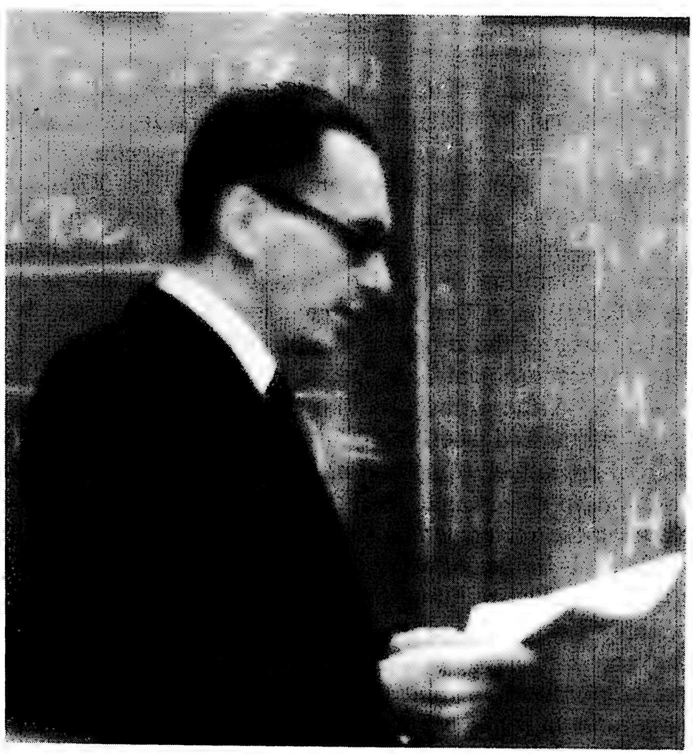
REFERENCES

1. Chung, F., Graham, R., Morrison, J. and Odlyzko, A., Pebbling a chessboard, *Amer. Math. Monthly* 102 (1995), 113–123.
2. Bishop, J. E., "Mathematicians find new key to old puzzle" *The Wall Street Journal*, Feb. 15, 1991 p. B1.
3. Bounds, D. G., "New optimization methods from physics and biology" *Nature* 329 (1987), 215–219.
4. Gardner, M., "The traveling salesman's travail," *Discover*, April 1985, pp. 87–90.
5. Knuth, D., "The traveling salesman problem," illustrative example in W. Sullivan, "Frontiers of Science, from Microcosm to Macrocosm," *The New York Times*, Feb. 24, 1976, p. 18.
6. Kolata, G., "Math problem, long baffling, slowly yields," *The New York Times*, Mar. 12, 1991, p. C1.
7. Lawler, E. L. et al. (Eds.), *The Traveling Salesman Problem*, Wiley & Sons, Chichester, 1985.
8. Port, O., "Solving the one about the traveling salesman," *Business Week*, June 2, 1986, p. 94.
9. Sangalli, A., "Short-circuiting the traveling salesman problem," *New Scientist*, June 27, 1992, p. 16.

Management & Data Systems
Lockheed Martin Corporation
King of Prussia, PA 19406

PICTURE PUZZLE

(from the collection of Paul Halmos)



A picture taken about a half century ago, at the Institute for Advanced Study.

(see page 329)

geometric discoveries were made possible through detailed constructions using geometric computer software packages.

ACKNOWLEDGMENTS. The author would like to record his thanks to Prof. H. S. M. Coxeter, Dr. T. J. Fletcher, Prof. H. B. Griffiths, and Dr. J. Rigby, together with the referee, for their helpful advice.

REFERENCES

1. Coxeter, H. S. M., *Introduction to Geometry*, Wiley, NY (1969), p. 15.
2. Kimberling, C., Central Points and Central Lines in the Plane of a Triangle, *Mathematics Magazine* Vol. 67, No. 3 (1994), pp. 163–187.
3. Soddy, F., The Kiss Precise, *Nature* 137 (1936), p. 1021.
4. Sommerville, D. M. Y., *Analytical Conics*, G. Bell, London (1924), p. 186.
5. Vandeghen, A., Soddy's Circles and the de Longchamps point of a triangle, this *Monthly* 71 (1964), 176–179.
6. Vandeghen, A., Some remarks on the isogonal and cevian transforms, this *Monthly* 72 (1965), 1091–1094.

Mathematics Centre
Chichester Institute of Higher Education
Bognor Regis, West Sussex, PO21 1HR
UNITED KINGDOM
a.oldknow@chihe.ac.uk

Finger-pointing in government has reached warp speed, traveling along a Moebius strip of infinite blame. Only courts can cut through it. Bringing back responsibility would stop the buck.

Philip K. Howard, *The Resurrection of Common Sense*,
The Wall Street Journal, Jan. 17, 1996, p. A18

Answer to Picture Puzzle

(page 307)

Kurt Gödel

Random Triangles in n Dimensions

Bennett Eisenberg and Rosemary Sullivan

1. INTRODUCTION. In 1886 the *Educational Times* published a collection of mathematical problems in which Woolhouse [**13**] proposes and solves the problem "If three points be taken at random in a given plane, prove that the probability of their being the vertices of an acute triangle is $4\pi^{-2} - 1/8$." His ingenious solution is correct except for the fact that he chooses the three vertices as independent, uniformly distributed points in the unit disc. Lewis Carroll in his *Pillow Problems* (1893) poses the problem of finding the probability that a "triangle formed by choosing three points at random on an infinite plane would have an obtuse angle." Guy [**7**] gives a variety of "solutions" to this problem, including Carroll's questionable solution. From this we see that a major difficulty consists in defining a random triangle. Criteria for a satisfactory definition are discussed in a recent article by Portnoy [**12**]. Rather than try to solve this conundrum, this article concentrates on some interesting aspects of the computation of the probabilities under a few simple definitions of random triangle. The first definition is an extension of that of Woolhouse.

Assume that the vertices of the triangle are independent and uniformly distributed in the unit ball in n dimensions (the *n-ball*). In this case Hall [**8**] finds integral formulas for the probability that the triangle is acute. For $n = 2$ he gets Woolhouse's result ($\approx .28$). For $n = 3$ he gets $33/70$ ($\approx .47$). Buchta [**3**] gives closed form expressions for the integrals given by Hall and shows how the probability that such triangles are acute increases as the dimension increases. He suggests without elaborating that the probabilities tell something about how the shape of the unit ball in n dimensions changes as n increases. We discuss the influence of the shape of the n-ball on the probabilities in Section 3.

Another geometric problem is to prove Hall's [**8**] conjecture, "that amongst convex domains in $\mathbb{R}^n$, the probability that three randomly chosen points form an acute triangle is maximized when the domain is the n-ball." A relevant result is in Langford [**10**], where direct computation seems to show that if three points are selected at random in a rectangle, then the probability of the triangle being acute is maximized when the rectangle is a square. We say "seems to show" because the formulas for the probabilities of being obtuse in terms of the ratio of the lengths of the sides of the rectangle are rather complicated, so that one must use a computer generated graph to draw this conclusion. The graph in the article is certainly convincing beyond a reasonable doubt, but a proof would be better. Langford's work solves a *Math Monthly* problem (Hawthorne [**9**]).

Of course, a random triangle whose vertices all lie in a unit disc is not the same as a triangle whose vertices are chosen at random in the plane. Indeed, one cannot define random variables that are uniformly distributed in the plane. That is, there is no probability measure on the plane that is translation invariant. One alternative is to choose the vertices with components that are independent standard normal random variables. In this way the distribution of the vertex will be rotationally

invariant if not translation invariant. We call a triangle whose vertices are independently chosen in this way a *Gaussian triangle*. Portnoy [12] gives a comparatively simple proof that the probability that a Gaussian random triangle in two dimensions is obtuse is $3/4$. Indeed, he finds that other natural definitions of a random triangle in two dimensions also lead to the probability $3/4$ of being obtuse and wonders if there is some underlying principle behind this.

In Section 2 we make use of the invariance of the Gaussian distribution under orthogonal transformations as well as the well known F-distribution from statistical theory to find the probability that a Gaussian random triangle in n dimensions is obtuse. The formulas in dimensions higher than two are not nearly as simple as the $3/4$ answer in two dimensions. We compare the numerical results for Gaussian random triangles to Hall and Buchta's results for random triangles in the unit ball in n dimensions. In this way we can see how much difference the definition of randomness makes in answering certain questions.

Instead of choosing three vertices of a triangle at random it might be more natural in some situations to have one of the vertices fixed. For example, we could consider the triangle formed by the center of a target and points where two darts land. We call these *pinned random triangles*. In Section 4 we compare results for pinned Gaussian triangles to those for triangles where one vertex is fixed at the origin and the other two vertices are independent and uniformly distributed in the unit ball. In so doing we find yet another situation where the probability that a random triangle in two dimensions is obtuse is $3/4$.

In Section 5 we return to two dimensions and study the distribution of the largest angle of a triangle with vertices chosen at random in the unit disc, thus extending the work of Woolhouse and Hall. A critical lemma was first used by Hall. He based it on Baddeley's [1] generalization of the Crofton differential equation, which was introduced by Crofton [5] in his treatment of probability in the 1885 edition of the Encyclopaedia Britannica. The title of Baddeley's article, "Integrals on a Moving Manifold and Geometric Probability" indicates how far geometric probability has come in 100 years. To keep the article self-contained we give an argument for the validity of Hall's result in two dimensions. We also take advantage of the computer package MAPLE, which was not available in 1982.

2. OBTUSE GAUSSIAN TRIANGLES. Let the vertices of a random triangle in n dimensions be $Q_1 = (X_1, \dots, X_n)$, $Q_2 = (Y_1, \dots, Y_n)$, $Q_3 = (Z_1, \dots, Z_n)$, where X_i, Y_i, Z_i for $i = 1$ to n are independent $N(0, 1)$ random variables. The joint density of Q_1 is then $(2\pi)^{-n/2} \exp(-\Sigma x_i^2/2)$, which is invariant under orthogonal transformations. The triangle $Q_1Q_2Q_3$ is then a Gaussian triangle.

Theorem 1. *The probability that a Gaussian triangle in n dimensions is obtuse is given by* $(3\Gamma(n)/(\Gamma(n/2))^2)\int_0^{1/3} x^{(n-2)/2}/(1 + x)^n \, dx$.

This probability is three times the probability that a random variable with a Snedecor F-distribution (see Bain [2], p. 221) with n degrees of freedom in the numerator and denominator takes a value less than $1/3$. In the case $n = 2$, the probability is easily seen to be $3/4$.

Proof: Since at most one angle of the triangle can be greater than 90°, the probability that the triangle is obtuse is 3 times the probability that $\angle Q_3Q_1Q_2$ is greater than 90°. Moreover, $\angle Q_3Q_1Q_2$ is greater than 90° precisely when the dot

product of the vectors $Q_2 - Q_1$ and $Q_3 - Q_1$ is negative. This dot product is $D = \sum_{i=1}^{n}(Y_i - X_i)(Z_i - X_i)$.

Consider the quadratic form $(Y - X)(Z - X)$, where X, Y, and Z are independent $N(0,1)$ random variables. This can be written $(X,Y,Z)A(X,Y,Z)^T$ where A is the matrix $\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 0 & 1/2 \\ -1/2 & 1/2 & 0 \end{pmatrix}$. This matrix has eigenvalues $3/2$, $-1/2$, and 0. Thus $A = KSK^T$, where K is an orthogonal matrix and S is a diagonal matrix with diagonal entries $3/2$, $-1/2$, and 0. If $(U,V,W) = (X,Y,Z)K$, then by the invariance of the Gaussian distribution under orthogonal transformations U, V, and W are also independent and $N(0,1)$. It follows that $(Y - X)(Z - X) = \frac{3}{2}U^2 - \frac{1}{2}V^2$, where U and V are independent $N(0,1)$ random variables. Thus $D = \sum_{i=1}^{n}\frac{3}{2}U_i^2 - \frac{1}{2}V_i^2 = \frac{3}{2}\chi_n^2 - \frac{1}{2}\bar{\chi}_n^2$, where χ_n^2 and $\bar{\chi}_n^2$ are independent chi-square random variables with n degrees of freedom. We then have $P(\angle Q_3Q_1Q_2 > 90°) = P(D < 0) = P(F < 1/3)$, where $F = \chi_n^2/\bar{\chi}_n^2$ has an F distribution with n degrees of freedom in the numerator and n degrees of freedom in the denominator. The theorem follows. ∎

If n is even there is an interesting formula for the integral in Theorem 1.

Corollary 1. *If $n = 2k$, where k is a positive integer, the probability that a Gaussian random triangle in n dimensions is obtuse is*

$$3\sum_{j=k}^{2k-1}\binom{2k-1}{j}\left(\frac{1}{4}\right)^{j}\left(\frac{3}{4}\right)^{2k-1-j}.$$

Proof: Making the substitution $y = (1 + x)^{-1}$ in the integral expression of Theorem 1 and simplifying shows that the desired probability equals

$$3\frac{\Gamma(2k)}{\Gamma(k)^2}\int_{3/4}^{1} y^{k-1}(1-y)^{k-1}\,dy.$$

This is just 3 times the probability that the median of a random sample of $2k - 1$ uniform random variables on $[0,1]$ takes a value greater than $3/4$ (see Bain [2], p. 160) and that is the same as 3 times the probability that at least k of these uniform random variables take values greater than $3/4$. The corollary follows. ∎

An elegant expression for the probability in Theorem 1, which is true for all n, but which is harder to evaluate, is given in the next corollary.

Corollary 2. *For any $n > 1$, the probability that a Gaussian random triangle is obtuse is*

$$\frac{3\Gamma(n)}{(\Gamma(n/2))^2 2^{n-1}}\int_0^{\pi/3}(\sin\theta)^{n-1}\,d\theta.$$

Proof: This follows by substituting $x = \tan^2(\theta/2)$ in the integral of Theorem 1 and then using the formula $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$. ∎

The following table compares results of Hall and Buchta with our own.

TABLE 1. Probability that a Random Triangle is Obtuse

Dimension	2	3	4	5	6	7	8	9	16	20
Unit Ball	.72	.53	.39	.29	.22	.17	.13	.09	—	—
Gaussian	.75	.59	.47	.38	.31	.26	.22	.18	.05	.03

We see that the probability of a triangle being obtuse appears to decrease to 0 in both cases as the dimension increases. Also the probability of being obtuse is smaller when the vertices are chosen uniformly in the unit ball than when they are chosen with normal distributions.

3. THE SHAPE OF THE n-BALL. Aside from symmetry, two asymptotic properties of the shape of the n-ball explain why random triangles in the n-ball tend to be acute for large n.

Property 1. For large n, most of the volume of the n-ball is near its surface.

Property 2. For large n, most of the volume of the n-ball is near the subspace of vectors with first coordinate 0.

These properties lead to the fact that a random triangle in the n-ball for large n will be approximately equilateral with high probability. Hall **[8]** attributes this observation to a personal communication from Bert Fristedt. It deserves more exposure. Here we give a formal proof of this fact.

In what follows $Q(n)$ is a random vector uniformly distributed in the n-ball and $|Q(n)|$ is its length. We next show that for large n, $Q(n)$ will be very close to the surface of the unit ball.

Lemma 1. *If $Q(n)$ is uniformly distributed in the n-ball, then $|Q(n)| \to 1$ in probability. That is, for any $\varepsilon > 0$,*

$$P(1 - |Q(n)| > \varepsilon) \to 0 \quad as\, n \to \infty.$$

Proof: If V_n is the volume of the unit ball in $\mathbb{R}^n$, then $t^n V_n$ is the volume of a ball of radius t in $\mathbb{R}^n$. Thus $P(|Q(n)| \leqq 1 - \varepsilon) = (1-\varepsilon)^n V_n / V_n = (1-\varepsilon)^n$, which converges to 0 for $\varepsilon > 0$. ■

We next show that if U_n is a unit vector in $\mathbb{R}^n$, then for large n, $Q(n)$ will be approximately orthogonal to U_n.

Lemma 2. *If for each n, U_n is a unit vector in $\mathbb{R}^n$ and $Q(n)$ is uniformly distributed in the n-ball, then $Q(n) \cdot U_n \to 0$ in probability.*

Proof: Write $Q(n)$ as $(X_1, \dots, X_n)$. By symmetry we may take U_n as $(1, 0, \dots, 0)$. Then $Q(n) \cdot U_n = X_1$. Now $E(X_1) = 0$ and $E(X_1^2) = E(|Q(n)|^2)/n < 1/n$, which approaches 0. Hence $Q(n) \cdot U_n$ converges to 0 in mean square and *a fortiori* in probability. ■

Similarly, two independently selected random vectors in the n-ball will be approximately orthonormal for large n.

Lemma 3. *If $Q_1(n)$ and $Q_2(n)$ are independent and uniformly distributed in the n-ball, then $Q_1(n) \cdot Q_2(n) \to 0$ in probability.*

Proof: Write $Q_1(n) = (X_1, \dots, X_n)$ and $Q_2(n) = (Y_1, \dots, Y_n)$. Then $E(Q_1 \cdot Q_2) = E(\sum_{i=1}^n X_i Y_i) = 0$ and $\mathrm{Var}(Q_1 \cdot Q_2) = E(\sum_{i=1}^n X_i^2 Y_i^2) = nE(X_1^2)E(Y_1^2) < 1/n$. The lemma follows. ■

We have shown that for large n, Q_1, Q_2, and Q_3 will be approximately orthonormal vectors. Thus the triangle determined by these three vertices will be approximately equilateral and will have very small probability of being obtuse.

Theorem 2. *Let Θ_n be the angle at $Q_1(n)$ in the triangle with independent vertices $Q_i(n)$ for $i = 1, 2, 3$. Then $\Theta_n \to \pi/3$ in probability.*

Proof: $\cos(\Theta_n) = (Q_2 - Q_1) \cdot (Q_3 - Q_1)/(|Q_2 - Q_1||Q_3 - Q_1|)$. We now use the fact (see Port **[11]**, p. 567) that if random variables converge in probability, then continuous functions of the random variables converge in probability to the corresponding limits. Here we have by Lemmas 1 and 3 that $(Q_2 - Q_1) \cdot (Q_3 - Q_1) \to 1$ in probability and $|Q_2 - Q_1|^2$ and $|Q_3 - Q_1|^2$ each converge to 2 in probability. Thus $\cos(\Theta_n) \to 1/2$ in probability. It follows that $\Theta_n \to \arccos(1/2) = \pi/3$ in probability. ■

In particular, we see that $P(\Theta_n \geqq \pi/2) \to 0$. Theorem 2 also can be proved when the components $(X_1, \dots, X_n)$ of $Q(n)$ are independent, mean 0 random variables with common variance σ^2 and common fourth moments. Normalized vectors $Q^*(n) = Q(n)/\sqrt{n}$ are needed to carry through the argument. We omit the details. We mention, however, that a careful analysis could perhaps be used to explain the difference in the rates of convergence of Θ_n as indicated by Table 1.

4. OBTUSE PINNED RANDOM TRIANGLES. There are other ways to define a random triangle than to choose the three vertices at random. One natural way is to start with one vertex fixed and to choose the other two vertices at random. We call such a random triangle a *pinned random triangle*. If we let $Q_1 = (0, \dots, 0)$, $Q_2 = (X_1, \dots, X_n)$, and $Q_3 = (Y_1, \dots, Y_n)$, where $X_1, \dots, X_n, Y_1, \dots, Y_n$ are independent $N(0,1)$ random variables, then the triangle $Q_1Q_2Q_3$ is called a *pinned Gaussian triangle*.

Theorem 3. *The probability that a pinned Gaussian triangle in n dimensions is obtuse is given by $1/2 + (2\Gamma(n)/(\Gamma(n/2))^2)\int_0^\alpha x^{(n-2)/2}/(1+x)^n\,dx$, where $\alpha = 3 - 2\sqrt{2}$.*

Proof: The proof is similar to that of Theorem 1. Consider first the dot product of $Q_2 - Q_1$ and $Q_3 - Q_1$. This is $\sum_{i=1}^n X_i Y_i$ and the probability that this is negative is 1/2 by symmetry. Thus $P(\angle Q_2Q_1Q_3 > 90°) = 1/2$.

The dot product of $Q_3 - Q_2$ and $Q_1 - Q_2$ is $\sum_{i=1}^n (Y_i - X_i)(-X_i)$. Since the eigenvalues of the matrix $A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 0 \end{pmatrix}$ are $\lambda_1 = (1 + \sqrt{2})/2$ and $\lambda_2 = (1 - \sqrt{2})/2$, we have as in the proof of Theorem 1 that $P(\angle Q_1Q_2Q_3 > 90°) = P(\lambda_1 \chi_n^2 + \lambda_2 \bar{\chi}_n^2 < 0) = P(F < 3 - 2\sqrt{2})$, where F has an $F(n,n)$ distribution. The same holds for $\angle Q_2Q_3Q_1$. The theorem follows. ■

It is interesting to compare the probabilities of being obtuse for a pinned Gaussian triangle and a *pinned random triangle in the n-ball*, i.e., a random triangle with one vertex at the origin and the other two vertices independent and uniformly distributed in the unit ball.

Theorem 4. *The probability that a pinned random triangle in the n-ball is obtuse is* $\frac{1}{2} + (\frac{1}{2})^n$.

Proof: Symmetry arguments tell us that we may assume that the random vertex that is further from the origin is at $(r, 0, \dots, 0)$, for some r, $0 < r < 1$, and the one closer to the origin is uniformly distributed in the ball of radius r. Call the closer point Q. For such a triangle the angle at the origin will be obtuse when the first coordinate of Q is negative and the angle at Q will be obtuse when it lies in the ball centered at $(r/2, 0, \dots, 0)$ with radius $r/2$. The angle at the vertex at $(r, 0, \dots, 0)$ must be acute. The shaded region in Figure 1 shows where Q must fall for the triangle to be obtuse for $n = 3$.

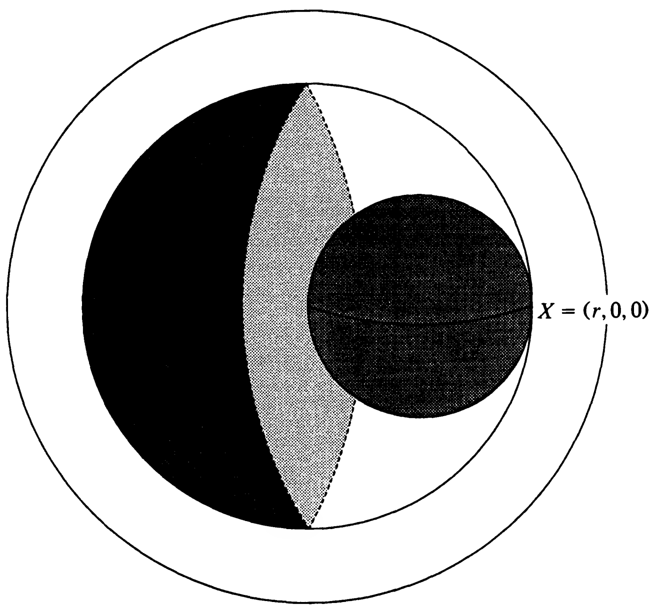


Figure 1. The set of Q where triangle OXQ is obtuse.

The probability that the first coordinate of Q is negative is $1/2$. Since the relative volume of a ball of radius $r/2$ to a ball of radius r in n dimensions is $(1/2)^n$, it follows that the probability that the angle at Q is obtuse is $(1/2)^n$. The theorem follows. ■

The following short table compares probabilities of being obtuse for pinned Gaussian triangles and pinned random triangles in the unit ball.

TABLE 2. Probability that a Pinned Triangle is Obtuse

Dimension	2	3	4	5	6	7	8
Unit Ball	.75	.63	.56	.53	.52	.51	.50
Gaussian	.79	.68	.62	.58	.55	.54	.52

In each case the probability of being obtuse decreases to .5. This equals the probability that the angle at the origin is obtuse and indicates that the probabilities that the other angles are obtuse are decreasing to zero. Indeed, one can see by the arguments of Section 3 that the angle at the origin will approach $\pi/2$ and the other angles will each approach $\pi/4$ as the dimension increases. Once again the probability of being obtuse is greater in the Gaussian case than in the unit ball case.

5. RANDOM TRIANGLES IN THE UNIT DISC. We now return to triangles in the plane, where more can be said. We restrict attention to the case where Q_1, Q_2, and Q_3 are uniformly distributed in the unit disc in $\mathbb{R}^2$. Let L be the largest angle in the triangle $Q_1Q_2Q_3$. Hall and Woolhouse show that $P(L \geqq \pi/2) = \frac{9}{8} - 4\pi^{-2} \approx .72$. In this section we extend Hall's argument in two dimensions to find $P(L \geqq t)$ for $t \geqq \pi/2$. Applying Baddeley's generalization of Crofton's theorem (Baddeley **[1]**), Hall shows that we may assume that two of the points lie on the circumference of a circle at an angle A with density $f(\alpha) = (1 - \cos\alpha)/\pi$ for $0 \leqq \alpha \leqq \pi$ and the third point is uniformly distributed inside the circle.

Why is this true? The argument begins like the one on pinned random triangles in the unit ball. Symmetry arguments and properties of the uniform distribution tell us that the point furthest from the origin may be assumed to be at $(r, 0)$ with the other two points uniformly distributed inside the circle of radius r centered at the origin. But the probability of the event of interest does not depend on r so we may assume $r = 1$. Now we move the circle for convenience. We place the fixed point at $(0, 0)$ and the random points inside the circle described in polar coordinates by $r = 2\cos(\theta)$. Each point then lies on one circle given by $r = 2t\cos(\theta)$.

Choose the point $Q = (X, Y)$ corresponding to the larger t. Let T be the random variable whose value is that t. Then the other point may be assumed to be uniformly distributed within the circle $r = 2T\cos(\theta)$. The only question is the distribution of the angle A between the point Q on the circle $r = 2T\cos(\theta)$, the center of that circle $(T, 0)$ and $(0, 0)$. Let Θ be the polar angle of Q, where $-\pi/2 \leqq \Theta \leqq \pi/2$, and let $B = 2\Theta$, the polar angle of Q where the origin is at $(T, 0)$ as in Figure 2. Then $A = \pi - |B|$.

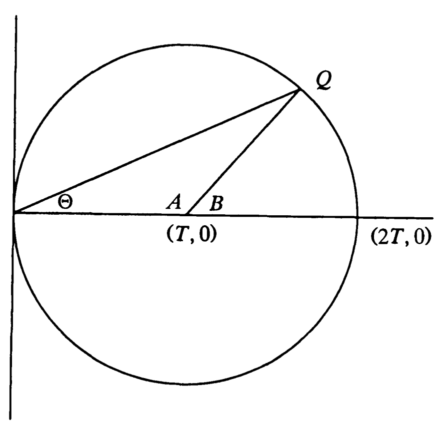


Figure 2. Angle A equals $\pi - |2\Theta|$.

To find the distribution of A, let (T_i, B_i) be the new coordinates of the point $Q_i = (X_i, Y_i)$, for $i = 1, 2$, where $X_i = T_i + T_i \cos(B_i)$ and $Y_i = T_i \sin(B_i)$. Applying the usual Jacobian rule for the transformation of random variables we find that (T_i, B_i) has density $f(t, \beta) = t(1 + \cos(\beta))/\pi$ for $0 \leqq t \leqq 1$ and $-\pi \leqq \beta \leqq \pi$. It follows that T_i and B_i are independent, where T_i has density $2t$ for $0 \leqq t \leqq 1$ and B_i has density $(1 + \cos(\beta))/2\pi$ for $-\pi \leqq \beta \leqq \pi$. Hence knowledge about T_i tells us nothing about B_i for $i = 1, 2$. Thus knowing that Q is equal to that Q_i corresponding to the larger T_i tells us nothing about the corresponding B_i. It follows that B for the point Q on the circle $r = 2T\cos(\theta)$ will have density $(1 + \cos \beta)/(2\pi)$. Since $A = \pi - |B|$, we get Hall's formula for the density of A.

For each $t \geq \pi/2$ we consider the conditional probability $P(L \geq t | A = \alpha)$. The calculation and formula depend on whether $\alpha > 2\pi - 2t$ or $\alpha < 2\pi - 2t$. In the case $t = \pi/2$ as treated by Hall, the division into two cases is unnecessary. We use the fact that for two given points Q_1 and Q_2 on the x axis and Q_3 in the upper half plane, the triangle $Q_1Q_2Q_3$ will have largest angle greater than t if and only if Q_3 lies in the shaded region in Figure 3.

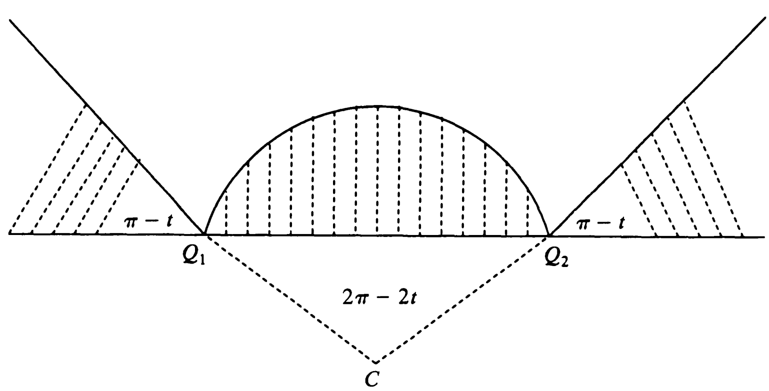


Figure 3. The set of Q where the largest angle of triangle QQ_1Q_2 is greater than t.

The shaded circular type region is based on the fact from plane geometry that if Q_3 lies on the circular arc centered at C with angle $2\pi - 2t$ at C, then $\angle Q_1Q_3Q_2 = t$. For clarity we assume that Q_3 must lie in the upper half plane. There is a symmetric diagram for Q_3 in the lower half plane.

It follows that the angle $\alpha = 2\pi - 2t$ is a critical value in computing probabilities. In this case the tangent line to the circle at Q_1 makes an angle of t with the secant from Q_1 to Q_2 (see Figure 4).

If $\alpha < 2\pi - 2t$ then the angle between the chord Q_1Q_2 and tangent line at Q_1 is greater than t so that the line at angle t with the chord at Q_1 lies inside the tangent line. Otherwise, it lies outside. It follows that for $\alpha > 2\pi - 2t$, the possible locations for Q_3 inside the circle where $L \geq t$ is given by the shaded region in Figure 5.

If $\alpha < 2\pi - 2t$ the diagram is more complicated. The possible locations where $L \geq t$ is given by the shaded region in Figure 6.

In either case the critical region for calculating probabilities is a disjoint union of segments of circles. Calculating the areas of the shaded regions in Figures 5 and 6 is easy, since the area of a segment of a circle of radius r is $H(r, \theta) = r^2(\theta - \sin \theta)/2$, where θ is the angle of arc of the segment.

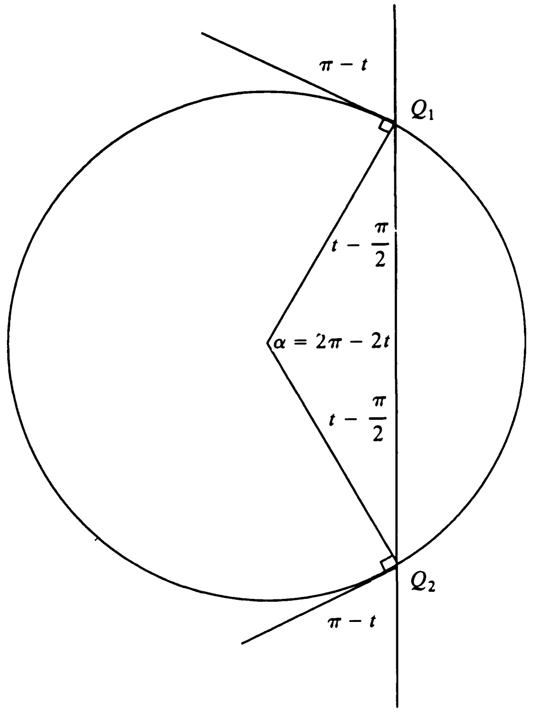


Figure 4. If $\alpha = 2\pi - 2t$, then the angle between the tangent line at Q_1 or Q_2 and the secant equals t.

Lemma 4. *If Q_1 and Q_2 lie on the circumference of a unit circle at an angle α apart and Q_3 is uniformly distributed inside the circle, then*

i) for $t \geqq \pi/2$ and $0 \leqq \alpha \leqq 2\pi - 2t$,

$$P(L \geqq t) = \left\{ \left(\frac{\sin^2(\alpha/2)}{\sin^2(t)} \right) \left[\pi - t + \frac{1}{2}\sin(2t) \right] - \frac{\alpha}{2} - \frac{\sin(\alpha)}{2} + 2\pi - 2t + \sin(2t + \alpha) \right\} / \pi$$

ii) for $t \geqq \pi/2$ and $2\pi - 2t \leqq \alpha \leqq \pi$,

$$P(L \geqq t) = 2\left(\frac{\sin^2(\alpha/2)}{\sin^2(t)} \right) \left[\pi - t + \frac{1}{2}\sin(2t) \right] / \pi$$

Proof: In case (i) Figure 6 applies and the shaded area is $2H(1, 2\pi - 2t - \alpha) + H(1, \alpha) + H(\sin(\alpha/2)/\sin(\pi - t), 2\pi - 2t)$.

In case (ii) Figure 5 applies and the shaded area is $2H(\sin(\alpha/2)/\sin(\pi - t), 2\pi - 2t)$. The result follows. ■

Combining Lemma 4 and Hall's formula for the density of A we have Theorem 5.

Theorem 5. *$P(L \geqq t) = \int_0^\pi P(L \geqq t | A = \alpha)(1 - \cos\alpha)/\pi \, d\alpha$, where $P(L \geqq t | A = \alpha)$ is given in Lemma 4.*

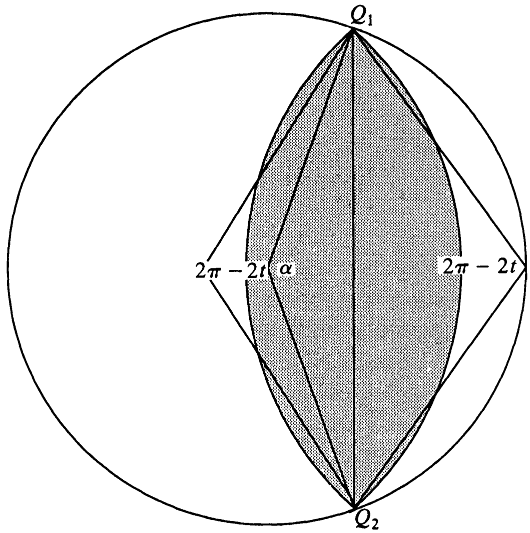


Figure 5. The region where $L > t$, if $\alpha > 2\pi - 2t$.

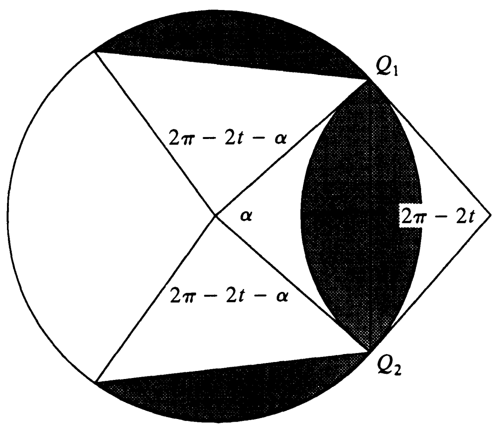


Figure 6. The region where $L > t$, if $\alpha < 2\pi - 2t$.

A closed form formula for $P(L \geqq t)$ can be found using MAPLE, but it is about a half page long and is not very illuminating. It does show that $P(L \leqq \pi/2) = 4\pi^{-2} - 1/8$ and $P(L \leqq 2\pi/3) = 27\pi^{-2}/8 + \sqrt{3}\,\pi^{-1}/12 + 2/9 = .61$. A short table of values of the cumulative distribution function $F(t) = P(L \leqq t)$ is given below.

TABLE 3. The Distribution of the Largest Angle of a Triangle in the Unit Disc

t	$.5\pi$	$.6\pi$	$.7\pi$	$.8\pi$	$.9\pi$	$.95\pi$	$.98\pi$	$.99\pi$	π
$P(L \leqq t)$	.28	.50	.66	.79	.90	.95	.98	.99	1.

It appears that as $t \to \pi$, $F(t) \sim t/\pi$. Indeed, MAPLE can formally compute the density $f(t)$ of L for $\pi/2 \leqq t \leqq \pi$. As expected, $f(\pi) = 1/\pi$. The table could have been found by numerical integration, but the exact values of $F(t)$ and $f(t)$ require computer algebra. It would be interesting to find a geometric explanation for the value of $f(\pi)$.

The main problem in extending these results to $t < \pi/2$ is the complexity of the analogous geometric diagram in Figure 3. In such a case more than one angle can be greater than t, making the calculation of critical areas virtually impossible. The results just for $t \geqq \pi/2$ are useful, however. The main statistical applications involve the distribution of L for large t. Here one tests whether the unusually flat shape of some triangle formed by three points occurring in nature is "random" or due to some cause. The significance level of the test is based on the distribution of L for large t.

The problems that motivated our original interest in random triangles can also be partly answered from this distribution for large t. One is the classical Fermat problem "to find the point in a triangle which minimizes the sum of the distances to the three vertices of the triangle." This point would determine the shortest path connecting the vertices. If the largest angle of the triangle is at least 120°, then the vertex of the largest angle is the solution to Fermat's problem (see Eisenberg and Khabbaz [6]). Otherwise, the point is inside the triangle. The question is, if the triangle is random, what is the probability that the "Fermat point" is a vertex of the triangle? The results above tell us that if the vertices of the triangle are chosen at random in the unit disc, then this probability is .39.

The other problem is related to that of finding the smallest circle containing a set of points, i.e., the minimal spanning circle. The question is, if the points are chosen at random, what is the probability that such a circle goes through exactly two of the points? In the case of three points this reduces to the probability that the three points form the vertices of an obtuse triangle (with minimal spanning circle having as diameter the longest side of the triangle). When the three points are chosen at random in the unit disc, the probability is approximately .72. If they are chosen with Gaussian distributions, the probability is $3/4$. How this probability behaves for a larger number of points is an open question.

REFERENCES

1. Baddeley, A., Integrals on a moving manifold and geometric probability. *Adv. Appl. Prob.* 9:588–603, 1977.
2. Bain, L. J. and M. Englehardt., *Introduction to Probability and Mathematical Statistics*. Duxbury Press, Boston, 1987.
3. Buchta, C., A Note on the Volume of a Random Polytope in a Tetrahedron. *Illinois J. Math.* 30:653–659, 1986.
4. Carroll, L., *Pillow Problems*, 1893. 4th edition (1895) Problem 58, pp. 14, 25, 83–84; reprinted (with A Tangled Tale), Dover Publications, New York, 1958.
5. Crofton, M. W., Probability. *Encyclopaedia Britannica*, Ninth edition, 19:768–788, 1885.
6. Eisenberg, B. and S. Khabbaz, Optimal Locations. *College Math. J.* 23:282–289, 1992.
7. Guy, R. K., There are three times as many obtuse-angled triangles as there are acute-angled ones. *Math. Mag.* 66:175–178, 1993.
8. Hall, G. R., Acute Triangles in the n-Ball. *J. Appl. Prob.* 19:712–715, 1982.
9. Hawthorne, F., Problem E1150. *Amer. Math. Month.* 62:40, 1955.
10. Langford, E., A Problem in Geometric Probability. *Math. Mag.* 43:237–244, 1970.
11. Port, S. C., *Theoretical Probability for Applications*, John Wiley and Sons, Inc., New York, 1994.
12. Portnoy, S., A Lewis Carroll Pillow Problem: Probability of an Obtuse Triangle. *Statist. Science* 9:279–284, 1994.
13. Woolhouse, W. S. B., Solution by the Proposer to Problem 1350. *Mathematical Questions, with their Solutions from the Educational Times, 1*, C. F. Hodgson and Son, London, 49–51, 1886.

Department of Mathematics
Lehigh University
Bethlehem, PA 18015
BE01@lehigh.edu

The Euler-Gergonne-Soddy Triangle of a Triangle

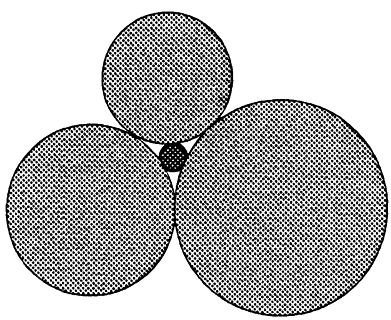
Adrian Oldknow

The circumcentre O, centroid G and orthocentre H of a triangle are collinear in the Euler line. However the incentre I lies on the Euler line only if the triangle is isosceles.

In general I belongs, together with three other centres Ge, S, S' (the Gergonne point and the inner- and outer-Soddy centres) to another important line identified some 30 years ago, and known as the Soddy line. In this article a number of new points on the Soddy line are identified as centres of perspective for pairs of triangles that have a common axis of perspective. That common axis, the Gergonne line, is shown to be orthogonal to the Soddy line and to contain other important points. Together with the Euler and Soddy lines, the Gergonne line forms the right-angled triangle that is the title of this article.

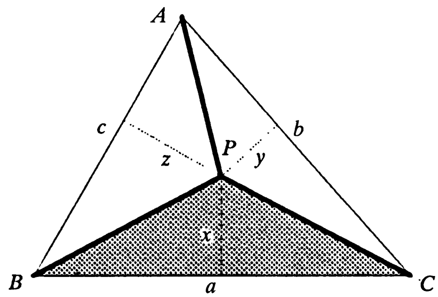
1. THE "FOUR-COINS" PROBLEM. The discoveries in this article arose from an innocuous-sounding problem. Take three coins (or circular disks) such as a penny, dime, and quarter. Arrange them so that each touches the other two. There is a small region between them bounded by three circular arcs. It looks as if a fourth small circle can be fitted into this so that it touches each of the three original circles.

How is its radius connected with those of the three circles?
Where is its centre in relation to the centres of the three circles?



Before discussing the solutions to these problems it may be helpful to summarize some of the geometric techniques and results that are useful in tackling such problems.

2. TRILINEAR COORDINATES. In this system each point P in the plane of the triangle of reference ABC is represented by the triplet (x, y, z) where x is the perpendicular distance from P to BC etc. x is taken as positive, zero, or negative depending on whether P lies on the same side of, on, or on the other side of the side BC from the vertex A.

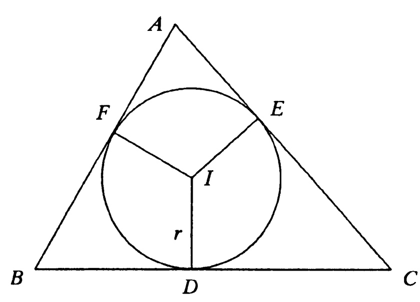


The values of x, y, z are not independent. If the lengths of the sides BC, CA, AB are denoted conventionally as a, b, c then the areas of the triangles BPC, CPA, APB, with this convention of signs, always sum to the area Δ of triangle ABC.

Thus:

$$ax + by + cz = 2\Delta. \tag{1}$$

If P has the exact trilinear coordinates (x, y, z) then any non-zero multiple (kx, ky, kz) can be taken as its homogeneous coordinates. If required, the value of k can be determined from (1) and the exact trilinears for P can be recovered.



For example the incentre I is the centre of the circle O_I that touches the sides BC, CA, AB at the points D, E, F. So $ID = IE = IF = r$, the in-radius. Thus the exact trilinears of I are (r, r, r), and homogeneous coordinates for I are $(1, 1, 1)$. Applying (1) we find $r(a + b + c) = 2\Delta$ and writing, conventionally, the perimeter $a + b + c = 2s$ we have an expression for r:

$$r = \frac{\Delta}{s}. \tag{2}$$

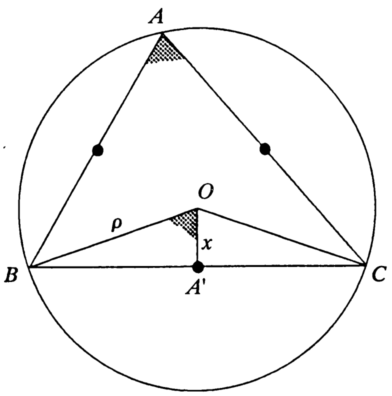
Note that this expression is *dimensionally correct*, in that Δ is an area, s is a length, and hence r is a length.

Provided that the reference triangle ABC is not degenerate then $\Delta > 0$ and so there is no real point in the plane with homogeneous coordinates that satisfy:

$$ax + by + cz = 0 \tag{3}$$

and this equation is known as that of *the line at infinity*.

3. THE EULER LINE. First it is necessary to find the coordinates of the circumcentre O and the orthocentre H. The detailed derivation will be given for O with that of H left as an exercise.



$AO = BO = CO = \rho$, the circumradius, and the angle at O subtended by the chord BC is twice the angle subtended at A. Hence the distance $x = OA'$ from O to BC is $x = \rho \cos A$.

Thus the exact trilinears for O are $(\rho \cos A, \rho \cos B, \rho \cos C)$, which give the homogeneous coordinates of O as $(\cos A, \cos B, \cos C)$. It is this simplicity of form, and the cyclic symmetry in the letters that is the power of trilinears (see [2]).

The homogeneous coordinates of H are $(\cos B \cos C, \cos C \cos A, \cos A \cos B)$. Provided the triangle is not right-angled this can be divided through by the factor $\cos A \cos B \cos C$ to give the equivalent form: $(\sec A, \sec B, \sec C)$.

The equation of a line joining two given points in trilinears is given by evaluating a simple determinant. Thus the equation of the Euler line OH is given by:

$$\begin{vmatrix} x & y & z \\ \cos A & \cos B & \cos C \\ \cos B \cos C & \cos C \cos A & \cos A \cos B \end{vmatrix} = 0$$

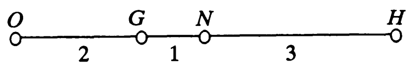
which simplifies to:

$$\sum x \cos A(\cos^2 B - \cos^2 C) = 0 \tag{4}$$

where the summation means the addition of two further terms formed by cycling the letters. Explicitly these are: $y \cos B(\cos^2 C - \cos^2 A)$ and $z \cos C(\cos^2 A - \cos^2 B)$.

The Euler line contains the centroid G, and the nine-point centre N (the circumcentre of A', B', and C'), with respective homogeneous coordinates (a, b, c) and $(\cos(B - C), \cos(C - A), \cos(A - B))$.

The four points $OGNH$ are collinear and lie with the fixed ratios: $OG : GN : NH = 2 : 1 : 3$.



Hence

$$OG : GN = 2 : 1 \quad \text{and} \quad OH : HN = 6 : 3 = 2 : 1.$$

So G and H divide O and N in the same ratio *internally* and *externally*. Such an arrangement of four collinear points $OGNH$ forms an *harmonic range*. Homogeneous coordinates are well suited for investigating harmonic ranges, although the precise values of particular ratios will be revealed only through the use of exact trilinears.

An alternative representation to (4) for the Euler line is the parametric form

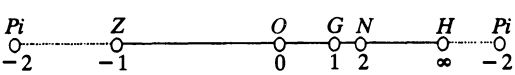
$$P(\lambda) = O + \lambda H. \tag{5}$$

If homogeneous coordinates are used for O and H, then the resulting function of λ gives the coordinates of the point P. As the parameter λ is varied, P sweeps out the Euler line. The circumcentre O is given by $\lambda = 0$ and the orthocentre H by $\lambda = \infty$. Similarly, the centroid G is at $\lambda = 1$ and the nine-point centre N at $\lambda = 2$.

If P and Q are two fixed points then for any non-zero parameter μ the four points P, $P + \mu Q$, Q, and $P - \mu Q$ form an harmonic range.

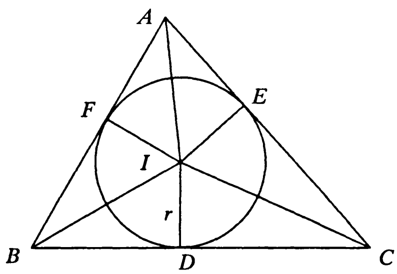
Using an alternative parametrization $P(\mu) = O + \mu N$ gives: $G = O + N$, $H = O - N$ and hence the four centres O, G, N, H form an harmonic range on the Euler line.

Note that the value of λ in $O + \lambda H$ that satisfies (3) is $\lambda = -2$ and hence the point $Pi = O - 2H$ is the *point at infinity* on the Euler line. The point given by the parameter $\lambda = -1$ is the de Longchamps point Z (see [5], [6]). It is such that O is the mid-point of ZH and has homogeneous coordinates: $(\cos A - \cos B \cos C, \cos B - \cos C \cos A, \cos C - \cos A \cos B)$.

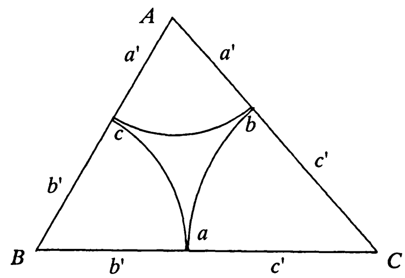


Note that, in general, the incentre I, with coordinates $(1, 1, 1)$ does not lie on the Euler line. It does so only if the triangle is isosceles, in which case the Euler line is the axis of symmetry. If the triangle is equilateral there is no unique axis of symmetry and the points O, G, N, H, Z all coincide with I. Pi becomes $(0, 0, 0)$ and the Euler line does not exist as a real line.

4. THE INCIRCLE O_I. The incircle O_I has centre I, radius r, and each of the sides BC, CA, AB as tangent. Let these sides touch the circle at D, E, F, respectively. Then, for example, $AE = AF = a'$ (say). The incentre I lies on the bisector of the angle at A, and this gives the classical construction for the point I as the intersection of the angle bisectors of the triangle. This also gives the key diagram for the "4-coin problem," and a', b', c' are the radii of the "coins" with centres A, B, C.



(a)



(b)

The perimeter

$$2s = a + b + c = (b' + c') + (c' + a') + (a' + b') = 2(a' + b' + c')$$

and hence

$$a' + b' + c' = s. \tag{6}$$

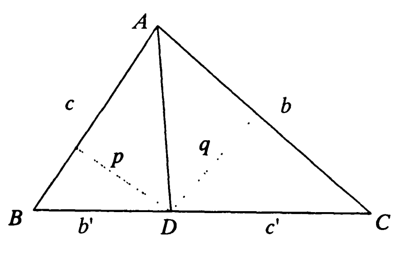
Also

$$a = b' + c' = a' + b' + c' - a' = s - a'$$

and hence

$$a' = s - a, \qquad b' = s - b, \qquad c' = s - c. \tag{7}$$

The trilinear coordinates of D, E, F are easily found. Suppose D has the exact coordinates $(0, p, q)$ and consider the area of the triangle ABD.



Thus: area(ABD) $= \frac{1}{2}pc = \Delta b'/a$ which gives $p = 2\Delta b'/ac$ and, similarly, $q = 2\Delta c'/ba$.

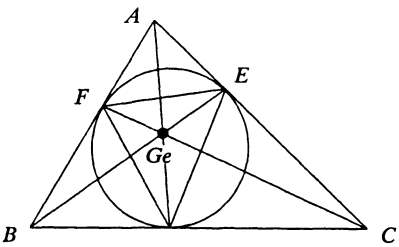
Writing

$$d = \frac{\Delta}{aa'}, \qquad e = \frac{\Delta}{bb'}, \qquad f = \frac{\Delta}{cc'} \tag{8}$$

gives the homogeneous coordinates of D,E,F as $(0, e, f), (d, 0, f), (d, e, 0)$.

The points D, E, F on the circumference of the incircle O_I of the triangle ABC, together with the vertices A, B, C define two related triangles. DEF will be called the *contact triangle* of ABC and ABC will be called the *tangent triangle* of DEF with respect to the given circle.

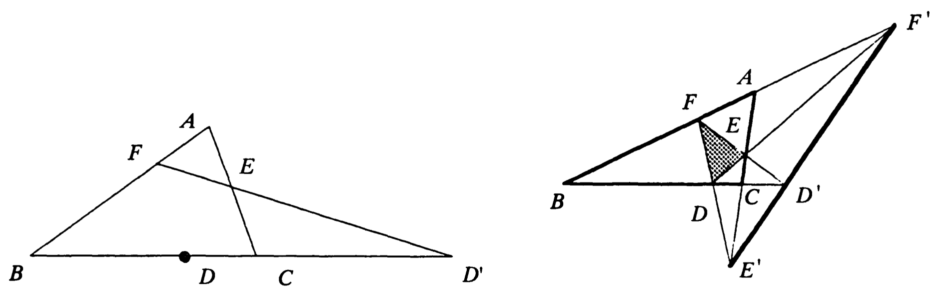
5. TRIANGLES IN PERSPECTIVE. The notion of *triangles in perspective* conjures up the image of a perspective drawing with a view-point (the *centre of perspective*), and a vanishing line or horizon (the *axis of perspective*).



Consider the triangles ABC and DEF above. The joins of the corresponding vertices AD, BE, CF meet in the Gergonne point Ge, which has coordinates (d, e, f). Thus ABC and DEF are a pair of triangles in perspective from the centre Ge.

In general, the contact and tangent triangles for any three points on a circle have a centre of perspective that does not coincide with the centre of the circle unless the points are equally spaced around the circle—in which case both triangles are equilateral.

6. THE GERGONNE LINE. From Desargues's theorem, a pair of triangles in perspective from a point (the centre) are also in perspective from a line (the axis). In the case of the incircle the tangent and contact triangles ABC, DEF are in perspective with centre Ge. Now consider the point D' where the corresponding sides BC and EF intersect.



(a) (b)

This point has coordinates $(0, e, -f)$, which can be easily verified by evaluating determinants to show that BCD' and EFD' are collinear sets. Using the paramet-

ric forms from §3:

$$D = B + \frac{f}{e}C, \qquad D' = B - \frac{f}{e}C$$

and hence $BDCD'$ form an harmonic range. If $b' = c'$ then $AB = AC$, so the triangle is isosceles and D is the mid-point of BC. EF is then parallel to BC and so D' becomes the point at infinity on BC.

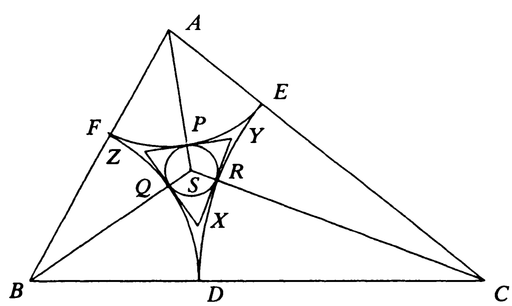
The three points $D'(0, e, -f), E'(-d, 0, f), F'(d, -e, 0)$ are collinear in the Gergonne line with equation:

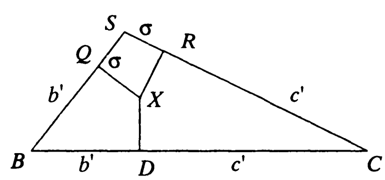
$$\frac{x}{d} + \frac{y}{e} + \frac{z}{f} = 0 \tag{9}$$

and this is the axis of perspective for the tangent and contact triangles ABC, DEF.

If $a' = b' = c'$ then the triangle is equilateral and $d = e = f$, in which case the equation of the Gergonne line becomes $x + y + z = 0$, which is the line at infinity. Note that the nomenclature "Gergonne line" is the author's own—the line plays an important role in the "4-coin problem."

7. THE SODDY LINE. Just as the Euler line has been defined parametrically as $O + \lambda H$ in terms of the circumcentre O and the orthocentre H so a line through the incentre I and the Gergonne point Ge can be defined as $I + \lambda Ge$. The point $S = I + Ge$, with $\lambda = 1$ is called the *inner-Soddy centre* of ABC. It is the solution to the second of the questions in §1. There is a circle O_S, with centre S and radius σ that touches each of the circles O_A, O_B, O_C (the three coins) externally. Suppose the points of contact are P, Q, R respectively.





Consider the triangle SBC. The circles O_S, O_B, O_C form another "3-coin" arrangement and so the common tangents to the circles at D, R, Q meet at X, the incentre of SBC. Hence X lies on the perpendicular to BC at D, which is the line ID. Also BX is the angle bisector of angle SBC. By similar arguments on triangle SAB the point Z is the intersection of the common tangents to O_S, O_A, O_B and hence lies on the line IF. Also BZ is the angle bisector of angle SBA. Hence the angle subtended by the tangent XZ at B is half the angle at B.

So the tangent triangle XYZ to the inner Soddy circle O_S for the contact triangle PQR:

(a) has its vertices X, Y, Z on the radii ID, IE, IF of the incircle O_I and
(b) has sides YZ, ZX, XY that subtend angles $\frac{1}{2}A, \frac{1}{2}B, \frac{1}{2}C$ at A, B, C,

and some rather involved trigonometry yields the additional result:

(c) $DX = d\,IX, \qquad EY = e\,IY, \qquad FZ = f\,IZ.$

Hence X, Y, Z have trilinears

$$(1, 1 + 2e, 1 + 2f), \qquad (1 + 2d, 1, 1 + 2f), \qquad (1 + 2d, 1 + 2e, 1),$$

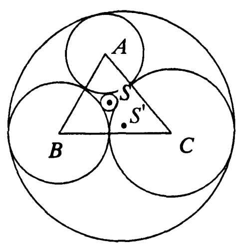
and those of P, Q, R are

$$(1 + 2d, 1 + e, 1 + f), \qquad (1 + d, 1 + 2e, 1 + f), \qquad (1 + d, 1 + e, 1 + 2f).$$

From these it is easy to derive $(1 + d, 1 + e, 1 + f) = I + G$ as the coordinates of S. The equation of the Soddy line can thus also be written as

$$\sum (f - e)x = 0 \tag{10}$$

8. THE SODDY CIRCLES $O_S, O_{S'}$. The English chemist Frederick Soddy (1877–1958) investigated the first of the two questions in §1. Starting with three "kissing" coins O_A, O_B, O_C there is a fourth, small, one O_S that just fits into the gap and "kisses" each of the three circles externally. Similarly there is a fifth, larger, circle $O_{S'}$ that surrounds and "kisses" each of the three coins internally. This is the *outer-Soddy circle*, which, together with the inner-Soddy circle, forms the "divorced pair" in Soddy's quaint jargon (see [3]).



Soddy found the radii σ, σ' of these circles. In the notation of this article these become

$$\frac{1}{\sigma} = \frac{2}{r} + \left(\frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'}\right), \qquad \frac{1}{\sigma'} = \frac{2}{r} - \left(\frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'}\right). \tag{11}$$

In fact these had already been discovered by Descartes in the 17th Century (see [1]).

As may be expected the point S' (the outer-Soddy centre) also has a simple trilinear representation as $S' = I - Ge$, and hence the four points $S'ISGe$ are collinear and form an harmonic range (see [5]).

9. THE 15 CENTRES OF PERSPECTIVE FOR O_I, O_S AND $O_{S'}$. It has already been shown that the in-circle O_I has a tangent triangle ABC and a contact triangle DEF, which are in perspective from the centre Ge on the Soddy line. We now have a corresponding pair of triangles XYZ and PQR for the inner-Soddy circle O_S, and another pair $X'Y'Z'$ and $P'Q'R'$ for the outer-Soddy circle $O_{S'}$. So each of these pairs must have a centre of perspective Ri and Ri', say, (the *inner-* and *outer-Rigby* points) and these have the form

$$Ri = I + \tfrac{4}{3}Ge, \qquad Ri' = I - \tfrac{4}{3}Ge.$$

Comparing with the reference triangle ABC we find that ABC, XYZ are in perspective from a centre Ol, and ABC, $X'Y'Z'$ from a centre Ol' (the *inner-* and *outer-Oldknow* points), which have the form

$$Ol = I + 2Ge, \qquad Ol' = I - 2Ge.$$

There are 15 ways of choosing any 2 from the 6 triangles and each such pair turns out to have a centre of perspective on the Soddy line given by the following table:

	DEF	PQR	XYZ	$P'Q'R'$	$X'Y'Z'$
ABC	Ge	S	Ol	S'	Ol'
DEF		Ol	I	Ol'	I
PQR			Ri	I	Gr
XYZ				Gr'	I
$P'Q'R'$					Ri'

There are just 10 distinct centres. The points Gr, Gr' (the *inner-* and *outer-Griffiths* points) have the form:

$$Gr = I + 4Ge, \qquad Gr' = I - 4Ge.$$

Thus the 10 collinear points: S', I, S, Ri, Ol, Gr, Ge, Gr', Ol', Ri'
have parameter values: $\lambda = -1, 0, 1, \frac{4}{3}, 2, 4, \infty, -4, -2, -\frac{4}{3}$.

By exhaustion there turn out to be 22 harmonic ranges, such as $S'ISGe$, among the 210 sets of 4 points chosen from these 10. (There are other harmonic ranges, too, such as $IXDX'$ etc.)

10. THE SINGLE AXIS OF PERSPECTIVE. From Desargues' theorem each of the 15 pairs of perspective triangles with centres on the Soddy line has a corresponding axis of perspective on which corresponding sides intersect. The Gergonne line is the axis for the triangles ABC, DEF. The intersections of the sides are at the *Nobbs* points D', E', F' (see §6).

Remarkably each of the sides BC, EF, QR, YZ, $Q'R'$ and $Y'Z'$ passes through D' (with similar results for E' and F'). Hence all 15 pairs of triangles are in perspective from the same axis—the Gergonne line $D'E'F'$.

Homogeneous coordinates are not, in general, well-suited for finding distances and angles, but there is a test for whether two lines are perpendicular (see [4]).

Two lines with equations $px + qy + rz = 0$ and $p'x + q'y + r'z = 0$ with respect to the triangle of reference ABC are perpendicular if and only if

$$pp' + qq' + rr' = (qr' + q'r)\cos A + (rp' + r'p)\cos B + (pq' + p'q)\cos C$$

Applying this to the equations (9) and (10) for the Gergonne and Soddy lines, we find that they are, indeed, orthogonal. They intersect at the *Fletcher* point Fl, which has the form

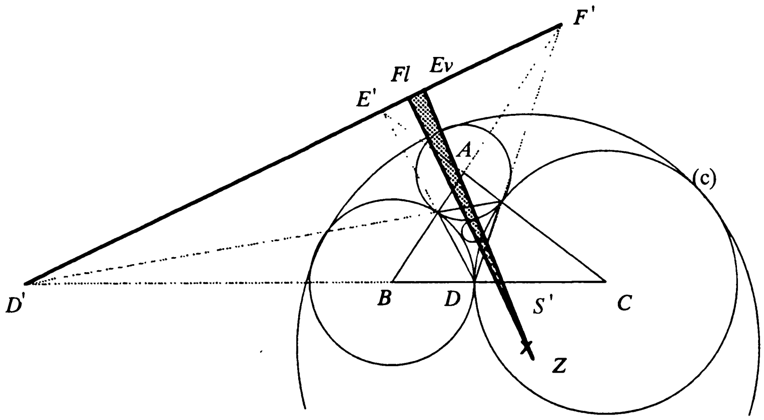
$$Fl = I - \frac{1}{3}\left(\frac{1}{d} + \frac{1}{e} + \frac{1}{f}\right)Ge.$$

Thus a general result has been established:

If D, E, F are any three points on a circle with centre I then the tangents at these points form another triangle ABC that is in perspective with DEF with centre Ge. The axis $D'E'F'$ is orthogonal to IGe and the point D' is the harmonic conjugate of D with respect to B, C etc.

11. THE EULER-GERGONNE-SODDY TRIANGLE. Thus:

(a) the circumcentre O and orthocentre H define the *Euler* line,
(b) the incentre I and the Gergonne point Ge define the *Soddy* line, and
(c) the Nobbs points D', E', F' define the *Gergonne* line.



Together, these three lines form a triangle $Fl Z Ev$, right-angled at the Fletcher point Fl. The intersection of the Soddy and Euler lines is the de Longchamps point Z. In terms of the Soddy line, the point Z has the form $Z = I - \lambda Ge$, where $\lambda = 2s\Delta/(abc + 2a'b'c')$. The remaining point Ev, the *Evans* point, is the intersection of the Gergonne and Euler lines, and does not appear to have a simple parametric form in terms of either line.

12. EXTENSIONS. The reader is invited to explore what happens in special cases, such as right-angled, isosceles, and equilateral triangles. It is also worth analyzing the circumcircle O_O of ABC. Let the tangents to O_O at A, B, C form the triangle $A'B'C'$. There must be a centre of perspective L for ABC, $A'B'C'$ together with an axis of perspective, which is perpendicular to OL.

Prove that

(a) L is the Lemoine point (a, b, c),
(b) the axis is the *Lemoine line* with equation $x/a + y/b + z/c = 0$, and
(c) this axis is perpendicular to OL, whose equation is $\Sigma x \sin(B - C) = 0$.

Are there any interesting points on OL? Do OL and the Lemoine line have any interesting relationships with the Euler, Gergonne, or Soddy lines? Investigate other circles connected with ABC such as the e-circles or the nine-point circle.

POSTSCRIPT. It seems, on first sight, remarkable that such beautiful properties of the Soddy and Gergonne lines escaped the notice of the many geometers of the past who have unveiled many more obscure results. Although the account given here uses relatively simple algebra and trigonometry, such results were established only after considerable manipulation aided by computer algebra. The actual

geometric discoveries were made possible through detailed constructions using geometric computer software packages.

ACKNOWLEDGMENTS. The author would like to record his thanks to Prof. H. S. M. Coxeter, Dr. T. J. Fletcher, Prof. H. B. Griffiths, and Dr. J. Rigby, together with the referee, for their helpful advice.

REFERENCES

1. Coxeter, H. S. M., *Introduction to Geometry*, Wiley, NY (1969), p. 15.
2. Kimberling, C., Central Points and Central Lines in the Plane of a Triangle, *Mathematics Magazine* Vol. 67, No. 3 (1994), pp. 163–187.
3. Soddy, F., The Kiss Precise, *Nature* 137 (1936), p. 1021.
4. Sommerville, D. M. Y., *Analytical Conics*, G. Bell, London (1924), p. 186.
5. Vandeghen, A., Soddy's Circles and the de Longchamps point of a triangle, this *Monthly* 71 (1964), 176–179.
6. Vandeghen, A., Some remarks on the isogonal and cevian transforms, this *Monthly* 72 (1965), 1091–1094.

Mathematics Centre
Chichester Institute of Higher Education
Bognor Regis, West Sussex, PO21 1HR
UNITED KINGDOM
a.oldknow@chihe.ac.uk

Finger-pointing in government has reached warp speed, traveling along a Moebius strip of infinite blame. Only courts can cut through it. Bringing back responsibility would stop the buck.

Philip K. Howard, *The Resurrection of Common Sense*,
The Wall Street Journal, Jan. 17, 1996, p. A18

Answer to Picture Puzzle

(page 307)

Kurt Gödel

Abstract Algebra Uses Homomorphisms

Saunders MacLane

When I first learned about abstract algebra from Emmy Noether in 1931—and from Van der Waerden's then new book on "Modern Algebra"—I came to realize that the construction of quotient groups is intimately connected with group homomorphisms (onto the quotient group). This was emphasized, for instance, in Noether's "first" and "second" homomorphism theorems.

Now I read in this Monthly about quotient groups—with no mention of homomorphisms—in a recent article "An Abstract Algebra Story" (vol. 102, March 1995, pp. 227–242) by U. Leron and Ed Dubinsky. The article is about teaching abstract algebra. At the start, the students have written programs with ISETL, to implement the group axioms—with examples of Z_{11} or Z_{12} with no mention of congruence or of groups of geometric symmetries. After some development of Lagrange's Theorem the students are given a program that multiplies cosets by elements and so are prodded into perhaps discovering that the suggested multiplication of cosets does satisfy the group axioms for the case of a normal subgroup. But homomorphisms are not named and the kernel of a homomorphism is not noted (although in Figure 5 there appears an equation that does amount to the assertion that a certain operator K is indeed a homomorphism). In other words, the leading idea of abstract algebra has been dropped.

There may be a psychological intent, said to be supported by an elaborate theoretical framework and research; it is averred that certain activities help the student to "construct" the mental processes, objects, and relations necessary for a meaningful understanding of the topics. In other words, this article subordinates mathematical content to purported pedagogical principles.

The full understanding of group theory, especially of homomorphisms and quotient groups, took a century to develop. We should not pretend that prepared programs on computers will do instead. Teaching, by text and talk (lectures) to convey ideas has been and will be the medium to convey hard-won ideas to new thinkers.

In this century, the teaching of algebra in colleges has advanced remarkably. In the beginning (1904) there were texts such as H. B. Fine's "A College Algebra," covering the binomial theorem, progressions, permutations, and probability. More advanced courses (in the 1930's) dealt with the Theory of Equations (J. M. Thomas, Louis Weisner, etc.) up through solutions of the cubic and the quartic, but with little real contact with Galois theory. Dickson's "Modern Algebraic Theories" appeared in 1926 and Albert's "Modern Higher Algebra" in 1937. By 1941 there were accessible texts in English on the new ideas of Modern Algebra and Galois Theory. In 1942 a good friend told me that "Survey of Modern Algebra" would not fly west of the Hudson river; by 1946, it did fly frequently. Similarly, in the 1960's the revival of research in group theory led to a greater emphasis on groups in algebra courses.

This century has seen a steady and effective advance in the sophistication and scope of the teaching of algebra in colleges—presenting new ideas with examples and conveying them to eager students. The Leron-Dubinsky approach is a step backwards.

Department of Mathematics
The University of Chicago
5734 University Avenue
Chicago, IL 60637

Response from Leron and Dubinsky. MacLane bases his attack on several alleged facts concerning our method. These "facts" are simply wrong: Our course does treat the fundamental homomorphism theorem extensively, it does present the geometric groups of symmetry, and it does *not* use "prepared programs." All this can be easily seen by glancing at our textbook referred to in the paper (especially Chapter 5 and pp. 49–52). The three isolated "scenarios" in our paper do not represent the entire content of the course; rather they are an illustration of an implementation.

As for using "prepared programs," the paper emphasizes many times (see especially the third "Idealized Reader" question on p. 230) that the whole idea is that students *construct* these programs for themselves, and in the process also gradually construct in their minds the desired mathematical concepts. Thus, MacLane is attacking straw men he himself has erected. We have no argument with the content that he (and we) cherish. We are trying to make that content more accessible to more people.

We believe that our paper contains enough argument and evidence to merit at least a serious debate. Unfortunately, MacLane's letter uses forceful statement rather than argument to express his opinion. The mathematical community has recently made encouraging progress towards trying to meet the needs of undergraduate students. MacLane's letter is a step backward.

Department of Education in Technology and Science
Technion-Israel Institute of Technology
Haifa 32000, ISRAEL
uril@techunix.technion.ac.il

Mathematics Department
Purdue University
West Lafayette, IN 47907-1395
bbf@cc.purdue.edu

The Mean Value Theorem is the midwife of calculus—not very important or glamorous by itself, but often helping to deliver other theorems that are of major significance.

E. J. Purcell and D. Varberg
Calculus with Analytic Geometry, Fifth Edition,
Prentice-Hall, Englewood Cliffs, N.J., 1987.

NOTES

Edited by: **John Duncan**

Roth's Removal Rule and the Rational Canonical Form

Robert E. Hartwig

One of the highlights in any advanced course on Linear Algebra is the derivation of the Rational Canonical Form of a matrix T over a general field $\mathbb{F}$. This canonical form states that there is an invertible matrix Q, so that

$$Q^{-1}TQ = \mathrm{diag}[L(\psi_1), L(\psi_2), \ldots, L(\psi_s)], \qquad (*)$$

where for any monic polynomial $f(\lambda) = f_0 + f_1\lambda + \cdots + f_n\lambda^n, (f_n = 1)$, $L(f)$ is its companion matrix. That is,

$$L = L[f(\lambda)] = \begin{bmatrix} 0 & \cdot & & \cdot & \cdot & -f_0 \\ 1 & 0 & & & & -f_1 \\ 0 & 1 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ & & & \cdot & & \\ \cdot & & & 1 & & \cdot \\ 0 & \cdot & \cdot & 0 & 1 & -f_{n-1} \end{bmatrix}.$$

The polynomials ψ_i are called the invariant factors of T, and interlace, i.e., $\psi_{i+1} | \psi_i$ for $i = s-1, \ldots, 1$.

The traditional proofs of this canonical form, however, are no trivial matter and usually use induction, quotient spaces, or modules and take several lectures **[1–4]**, **[6]**.

In this note we present a "pure" matrix proof, which can be done in half of a class period, or less. It is based on Roth's Removal Rule **[5]** or **[4]** p. 422, which states that if $AX - XB = C$, then

$$\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}.$$

In terms of subspaces, this identity shows that in order to find a complementary invariant subspace it suffices to solve the associated Sylvester matrix equation. This we shall do explicitly for the case where B is a companion matrix.

Throughout this note $R(\cdot)$ denotes the range of $(\cdot)$.

Our first step shall be to associate with the given polynomial $f(\lambda)$, its adjoint polynomials $f_k(\lambda) = f_{k+1} + f_{k+2}\lambda + \cdots + \lambda^{n-k+1}$, $k = -1, 0, \ldots, n$ and its Han-

kel matrix

$$G = \begin{bmatrix} f_1 & f_2 & \cdots & & f_n \\ f_2 & f_3 & \cdots & f_n & \\ \vdots & & & \cdot & \\ \vdots & & \cdot & & \\ f_n & \cdot & & 0 & \end{bmatrix}, \qquad f_n = 1.$$

It is easily seen that $L^k \underline{e}_1 = \underline{e}_{k+1}$ for $k = 0, 1, \ldots, n-1$, and that $LG = GL^T$. That is, $G^{-1}LG = L^T$.

Next we observe that if $A \in \mathbb{F}_{m \times m}$, $L = L[f(\lambda)]$, $C = [\underline{c}_1, \ldots, \underline{c}_n]$ and $M = \begin{bmatrix} A & C \\ 0 & L \end{bmatrix}$ is $N \times N$ $(N = m + n)$, then it swiftly follows that

$$M^k = \begin{bmatrix} A^k & \Gamma(\lambda_k^k) \\ 0 & L^k \end{bmatrix},$$

where $\Gamma(\lambda^k) = A^{k-1}C + A^{k-2}CL + \cdots + CL^{k-1}$. Consequently

$$f(M) = \begin{bmatrix} f(A) & \Gamma(f) \\ 0 & 0 \end{bmatrix},$$

in which $\Gamma(f) = \sum_{j=0}^{n-1} f_j(A)CL^j$, and $0 = f(L)$.

To find a solution to $AX - XL(f) = C$ we equate columns and solve the recurrence relation. This leads us to define

$$\underline{\beta} = \Gamma(f)\underline{e}_1 = \sum_{j=0}^{n-1} f_j(A)CL^j\underline{e}_1 = \sum_{j=0}^{n-1} f_j(A)\underline{c}_{j+1}$$

and

$$E = \left[\underline{0}, \underline{c}_1, A\underline{c}_1 + \underline{c}_2, A^2\underline{c}_1 + A\underline{c}_2 + \underline{c}_3, \ldots, A^{n-2}\underline{c}_1 + A^{n-3}\underline{c}_2 + \cdots + \underline{c}_{n-1}\right].$$

Needless to say, if $f(M) = 0$ then $\Gamma(f) = 0$ and hence $\underline{\beta} = \underline{0}$. Now by direct computation it is easily seen that

$$A(-E) - (-E)L = C + \left[\underline{0}, \ldots, \underline{0}, \underline{\beta}\right].$$

Hence if $f(M) = 0$, then $\underline{\beta} = \underline{0}$, which ensures that the matrix equation $AX - XL = C$ has a solution $-E$. It now follows using Roth's Removal Rule, that

$$\begin{bmatrix} I & E \\ 0 & I \end{bmatrix} \begin{bmatrix} A & C \\ 0 & L \end{bmatrix} \begin{bmatrix} I & -E \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & L \end{bmatrix}. \qquad (**)$$

That is, $\begin{bmatrix} A & C \\ 0 & L \end{bmatrix}$ *is similar* to $\begin{bmatrix} A & 0 \\ 0 & L \end{bmatrix}$. All that now remains is to start the reduction procedure and put the pieces together.

Suppose that T is an $N \times N$ matrix $(N = m + n)$ with minimal polynomial $\psi_T = f(\lambda) = p_1^{m_1} p_2^{m_2} \cdots p_s^{m_s}$, of degree n, with p_i prime. If we select any $\underline{x}_i \in \{N[p_i^{m_i}(T)] \setminus N[p_i^{m_i - 1}(T)]\}$ and set $\underline{x} = \underline{x}_1 + \cdots + \underline{x}_s$, then it is routine to show (see the Appendix) that the minimal polynomial of the vector $\underline{x}$ is precisely $\psi_{\underline{x}}(\lambda) = f(\lambda)$. Its degree is maximal among all such polynomials. Next we complete the chain $[\underline{x}, T\underline{x}, \ldots, T^{n-1}\underline{x}]$ in any way to a basis matrix Q for $\mathbb{F}^N$. Then $Q^{-1}TQ = M = \begin{bmatrix} L & B \\ 0 & D \end{bmatrix}$, where $L = L(f)$ and $\psi_D | f$. The concluding step is to move the companion matrix to its "proper" position. In fact, it is straightforward to verify

that

$$\begin{bmatrix} L & B \\ 0 & D \end{bmatrix} \approx \begin{bmatrix} L^T & G^{-1}B \\ 0 & D \end{bmatrix} \approx \begin{bmatrix} D & 0 \\ G^{-1}B & L^T \end{bmatrix} = \begin{bmatrix} D^T & B^TG^{-1} \\ 0 & L \end{bmatrix}^T = \begin{bmatrix} A & C \\ 0 & L \end{bmatrix}^T$$

and

$$\begin{bmatrix} A & 0 \\ 0 & L \end{bmatrix}^T = \begin{bmatrix} D^T & 0 \\ 0 & L \end{bmatrix}^T = \begin{bmatrix} D & 0 \\ 0 & L^T \end{bmatrix} \approx \begin{bmatrix} L^T & 0 \\ 0 & D \end{bmatrix} \approx \begin{bmatrix} L & 0 \\ 0 & D \end{bmatrix}.$$

Transposing $(**)$ now yields $\begin{bmatrix} L & B \\ 0 & D \end{bmatrix} \approx \begin{bmatrix} L & 0 \\ 0 & D \end{bmatrix}$ as desired. It goes without saying that we may repeat the above steps with D to obtain the Rational Canonical Form $(*)$ of T.

The uniqueness of this form follows at once, if we recall that $\psi_1 = \psi$ is unique, and then apply the elementary result that

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \approx \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} \Rightarrow \psi_B = \psi_C.$$

In looking over the proof we see that Roth's Removal rule was crucial in the above derivation. Actually only half of this rule was used. It is of interest to note (and perhaps of classroom use) that, analogously to the above, we may prove the following version of Roth's theorem.

Theorem. *Let $L = L(f) \in \mathbb{F}_{n\times n}$ and $A \in \mathbb{F}_{m\times m}$. Then the following are equivalent:*

(i) *$AX - XL = C$ has a solution,*
(ii) $M = \begin{bmatrix} A & C \\ 0 & L \end{bmatrix} \approx K = \begin{bmatrix} A & 0 \\ 0 & L \end{bmatrix}$,
(iii) $R[\Gamma(f)] \subseteq R[f(A)]$, *and*
(iv) $\underline{\beta} = \Gamma(f)\underline{e}_1 \in R[f(A)]$.

In particular, if $f(M) = 0$ then $f(M)\underline{e}_{m+1} = 0$ and $\underline{\beta} = \Gamma(f)\underline{e}_1 = \underline{0}$, in which case a particular solution is given by $X = -E$.

Proof: (i) $\Rightarrow$ (ii). This is Roth's Removal Rule via $\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$.
(ii) $\Rightarrow$ (iii). If $MQ = QK$ then $\begin{bmatrix} f(A) & \Gamma(f) \\ 0 & 0 \end{bmatrix} Q = f(M)Q = Qf(N) = Q\begin{bmatrix} f(A) & 0 \\ 0 & 0 \end{bmatrix}$. Hence rank $[f(A), \Gamma(f)] = \text{rank}[f(A)]$, ensuring that $R[\Gamma(f)] \subseteq R[f(A)]$.
(iii) $\Rightarrow$ (iv). Clear.
(iv) $\Rightarrow$ (i). Let $f(A)\underline{s} = \underline{\beta}$ and $\mathscr{C} = [\underline{s}, A\underline{s}, \ldots, A^{n-1}\underline{s}]$ and define E as above. Then $X = \mathscr{C} - E$ is a solution to $AX - XD = C$, because

$$A(-E) - (-E)L = C + [0, \underline{\beta}] \qquad \text{and} \qquad A\mathscr{C} - \mathscr{C}L = [0, f(A)\underline{s}].$$

Moreover, if $f(M) = 0$ then $\Gamma(f)$ must vanish and so does $\underline{\beta} = \Gamma(f)\underline{e}_1$. We note in passing that all we really need is that $f(M)\underline{e}_{m+1} = 0$.

Appendix

If $\psi_T = f(\lambda) = p_1^{m_1}p_2^{m_2} \cdots p_s^{m_s} = \Psi_i(\lambda p_i^{m_i}(\lambda)$ then $N[p_i^{m_i}(T)] \neq N[p_i^{m_i-1}(T)]$. Otherwise $p_i^{m_i}(T)Y = p_i^{m_i-1}(T)$ for some Y and thus $0 = \Psi_i(T)p_i^{m_i}(T)Y = \Psi_i(T)p_i^{m_i-1}(T)$ implying that $\psi_T | \Psi_i(\lambda) \cdot p_i^{m_i-1}(\lambda)$, which is impossible. Next let

$\underline{x} = \underline{x}_1 + \cdots + \underline{x}_s$, with $\psi_{\underline{x}}(\lambda) = \phi(\lambda)$. Then $0 = \Psi_i(T)\phi(T)\underline{x} = \phi(T)\Psi_i(T)\underline{x} = \phi(T)\Psi_i(T)\underline{x}_i$. Thus $p_i^{m_i} | \phi(\lambda)\Psi_i(\lambda) \Rightarrow p_i^{m_i} | \phi(\lambda) \Rightarrow \psi | \phi$. The converse is clear, ensuring that $\underline{x}$ belongs to ψ_T.

REFERENCES

1. Herstein, I. N., *Topics in Algebra*, Ginn and Company, Waltham, 1964, p. 162.
2. Hoffman, K. and Kunze, K., *Linear Algebra*, lst ed., Prentice Hall, Englewood Cliffs, N.J., 1961.
3. Gantmacher, F. R., *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1960.
4. Lancaster, P. and Tismenetsky, M., *The theory of matrices*, second edition, Academic Press, 1985.
5. Roth, W. E., *The Equations $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc. 3:392–396 (1952).
6. Turnbull, H. W. and Aitken, A. C., *An Introduction to the theory of Canonical Matrices*, Second Impression, Blackie and Sons, 1945.

Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205

A Simple Proof of the Gale-Ryser Theorem

Manfred Krause

We consider matrices A of zeros and ones and shall be interested in the sum $c(A)$ of their column vectors and the sum $r(A)$ of their row vectors. Given *compositions* $(p_1, \ldots, p_k), (q_1, \ldots, q_l)$ of a positive integer n, i.e., non-negative integers $p_1, \ldots, p_k, q_1, \ldots, q_l$ such that $p_1 + \cdots + p_k = n = q_1 + \cdots + q_l$, does there exist a $k \times l$ matrix A of zeros and ones such that $c(A) = (p_1, \ldots, p_k), r(A) = (q_1, \ldots, q_l)$? An elegant answer to this combinatorial question is given by the important criterion by Gale and Ryser [**5**, chapter 6, theorem 1.1], which plays a prominent role in various mathematical areas and is commonly viewed as a fairly intricate result. For example, in [**2**, 1.4] it is obtained as a consequence of substantial parts of representation theory, in [**1**, 6.2.4] it is derived (in a generalized form) by means of graph theoretical methods, while [**4**, I.6, Example 2] refers to the proof in [**5**]. The object of this note is to propose a straightforward line of reasoning for the Gale-Ryser criterion.

It is easily seen that it suffices to consider the case that $p = (p_1, \ldots, p_k), q = (q_1, \ldots, q_l)$ are *partitions* of a positive integer n, i.e., that

$$p_1 \geq \cdots \geq p_k > 0, q_1 \geq \cdots \geq q_l > 0, \text{ and } p_1 + \cdots + p_k = n = q_1 + \cdots + q_l.$$

For example, $(3, 2, 2, 2, 1)$ and $(3, 3, 3, 1)$ are partitions of 10.

Any partition $p = (p_1, \ldots, p_k)$ may be visualized by a $k \times p_1$ matrix $A = (a_{i,j})$ of zeros and ones called the *Ferrers matrix* for p and defined by $c(A) = p$ and the following property: If $a_{i,j} = 0$, then $a_{i,k} = 0$ for all $k \geq j$. By transposing A we obtain again a Ferrers matrix A^*, and $p^* := c(A^*)$ is called the *conjugate* partition

of p. For example,

$$(3,2,2,2,1) \sim \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad (3,2,2,2,1)^* \sim \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

hence $(3,2,2,2,1)^* = (5,4,1)$. More formally, $p^* = (p_1^*, \ldots, p_{p_1}^*)$ where

$$p_i^* := |\{j\colon 1 \leq j \leq k,\, p_j \geq i\}| \quad \text{for all } i \leq p_1.$$

Given arbitrary compositions $p = (p_1, \ldots, p_k), q = (q_1, \ldots, q_l)$ of a positive integer n, we say that q is *dominated* by p if $\Sigma_{i=1}^m q_i \leq \Sigma_{i=1}^m p_i$ for all positive integers m, where $q_i := 0$ for all $i > l$ and $p_i := 0$ for all $i > k$. In this case we write $p \trianglerighteq q$. For example $(3,3,3,1)$ is dominated by $(5,4,1)$, but there is no dominance relationship between $(3,1,3,3)$ and $(3,2,2,2,1)$.

Theorem (Gale-Ryser). *Let p, q be partitions of a positive integer. Then there exists a $(0,1)$-matrix A such that $c(A) = p$, $r(A) = q$ if and only if q is dominated by p^*.*

The **necessity** of the dominance condition is commonly considered as the trivial part of the theorem [**3**, 4.3.19]. For the reader's convenience we recall a line of reasoning for it where we assume, more generally, that q is a composition while p is a partition. Let A be a $k \times l$ matrix of zeros and ones such that $c(A) = p$ and $r(A) = q$. If A does not contain *gaps*, i.e., if there are no $i \leq k$, $j < h \leq l$ such that $a_{i,j} = 0$ and $a_{i,h} = 1$, then A is a Ferrers matrix and $p^* = q$. Now let (i, j) be a gap of A and let $h > j$ be maximal such that $a_{i,h} = 1$. Swapping $a_{i,j}$ and $a_{i,h}$ we obtain a matrix $\tilde{A}$ such that $c(\tilde{A}) = p$, and the number of gaps of $\tilde{A}$ is lower than the number of gaps of A. As $r(\tilde{A}) \trianglerighteq r(A)$, it follows by induction on the number of gaps that $p^* \trianglerighteq q$, as asserted.

In the sequel we prove the **sufficiency** of the dominance condition for the existence of a matrix A such that $c(A) = p, r(A) = q$. Let $p = (p_1, \ldots, p_k), q = (q_1, \ldots, q_l)$ be partitions of a positive integer n such that $p^* \trianglerighteq q$. First we observe that there is a $k \times l$ matrix B of zeros and ones such that $c(B) = p$ and $r(B) \trianglerighteq q$, namely, the Ferrers matrix for p with $l - p_1$ columns of zeros adjoined. Thus, it will suffice to prove the following

Claim. Given a $k \times l$ matrix A of zeros and ones such that $c(A) = p$, $r(A) \trianglerighteq q$ and $r(A) \neq q$, we can find a $k \times l$ matrix A' of zeros and ones such that $c(A') = p$, $r(A') \trianglerighteq q$, and $\|r(A') - q\| < \|r(A) - q\|$ (where $\| \ \|$ is the ordinary Euclidean norm).

Since $\|r(A) - q\|^2$ is an integer, after a finite number of steps we are done.

Proof: Let $r(A) = (r_1, \ldots, r_l)$. Let i be minimal such that $r_i > q_i$, and let j be minimal such that $r_j < q_j$. Then $i < j$, since $r(A) \trianglerighteq q$. Clearly, for the vector

$$r' := (r_1', \ldots, r_l') := (r_1, \ldots, r_{i-1}, r_i - 1, r_{i+1}, \ldots, r_{j-1}, r_j + 1, r_{j+1}, \ldots, r_l)$$

we have $\|r' - q\| < \|r - q\|$. Moreover, q is dominated by r', as $r'_s \geq q_s$ for all $s < j$ and $r'_1 + \cdots + r'_s = r_1 + \cdots + r_s \geq q_1 + \cdots + q_s$ for all $s \geq j$. Since

$$r_i > q_i \geq q_j > r_j,$$

we can find a row index h (in fact there are at least two choices for h) such that $a_{h,i} = 1$ and $a_{h,j} = 0$. For any such h, the matrix $A' = (a'_{s,t})$ defined by swapping $a_{h,i}$ and $a_{h,j}$, i.e.,

$$a'_{s,t} := \begin{cases} 1 & \text{if } (s,t) = (h,j) \\ 0 & \text{if } (s,t) = (h,i) \\ a_{s,t} & \text{otherwise} \end{cases},$$

has row vector sum r' and column vector sum p. ∎

For example, by the theorem there must exist a matrix A such that $c(A) = (3,2,2,2,1)$, $r(A) = (3,3,3,1)$ because $(3,2,2,2,1)^* = (5,4,1) \trianglerighteq (3,3,3,1)$. Starting with the extended Ferrers matrix B for $(3,2,2,2,1)$, the procedure obtained from our proof leads to a solution for A as follows:

$$B = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\to \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = A.$$

By modifying A appropriately, (for example, by permuting the first three columns of A), the reader will find many other possibilities in this case. Note that the steps of the procedure are not uniquely determined. In particular the proof shows that in the case of $p^* \neq q$ we can always find at least two matrices A such that $c(A) = p$ and $r(A) = q$. Finally, we remark that the generalized theorem [**1**, 6.2.4] may easily be obtained along the same lines.

We should like to thank the referee for his useful suggestions.

REFERENCES

1. Brualdi, R., Ryser, H. J., *Combinatorial Matrix Theory*, Cambridge, New York, Port Chester, 1991.
2. James, G., Kerber, A., *The Representation Theory of the Symmetric Group*, Reading, Massachusetts, 1981.
3. Kerber, A., *Algebraic Combinatorics via Finite Group Actions*, Mannheim, Wien, Zürich, 1991.
4. Macdonald, I. G., *Symmetric functions and Hall polynomials*, Oxford, 1979.
5. Ryser, H. J., *Combinatorial Mathematics*, Wiley, 1963.

Mathematisches Seminar der Universität Kiel
Ludewig-Meyn-Str. 4
D-24098 Kiel
GERMANY

A Note on the Tennis Ball Theorem

Yves Martinez-Maure

The conclusion of the Tennis Ball Theorem (resp. the Four Vertex Theorem) may be reinforced under more restrictive assumptions. The purpose of this Note is to present such a result by dealing with envelopes parametrized by their Gauss map.

In this Note we are interested in the following type of plane curve.

Definition. *For any $h \in \mathscr{C}^{\infty}(S^1; R)$ we let $\mathscr{H}_h$ denote the envelope of the family of lines given by*

$$x \cos \theta + y \sin \theta = p(\theta), \tag{1}$$

where $p(\theta) = h(\cos \theta, \sin \theta)$.

We say that $\mathscr{H}_h$ is the hedgehog defined by the support function h, and that $\mathscr{H}_h$ is projective if p is a Möbius function (i.e., a 2π-periodic function such that $p(\theta + \pi) = -p(\theta)$ for all θ).

Figure 1 shows some examples of hedgehogs. A study of hedgehogs (envelopes parametrized by their Gauss map) in higher dimensions is given by R. Langevin, G. Levitt, and H. Rosenberg [2].

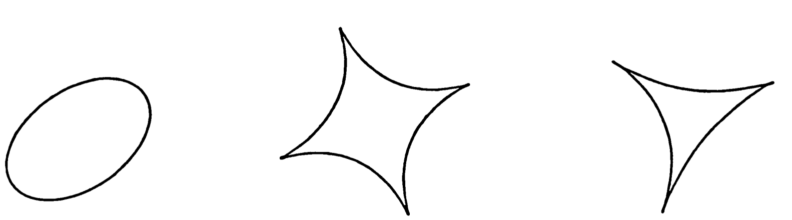
Partial differentiation of (1) yields

$$-x \sin \theta + y \cos \theta = p'(\theta). \tag{2}$$

From (1) and (2), the parametric equations for $\mathscr{H}_h$ are

$$\begin{cases} x = p(\theta)\cos \theta + p'(\theta)\sin \theta \\ y = p(\theta)\sin \theta + p'(\theta)\cos \theta \end{cases} \tag{3}$$

Suppose that $\mathscr{H}_h$ has a well-defined tangent line at the point (x, y), say T. Then T can be expressed by (1): the unit vector $u(\theta) = (\cos \theta, \sin \theta)$ is normal to T and $p(\theta)$ may be interpreted as the signed distance from the origin to T. Thus a singularity-free plane hedgehog is simply a convex curve (see Figure 1a). Furthermore, a not too singular plane hedgehog (i.e., a plane hedgehog that has a well-defined tangent line at every point) is a curve that has exactly one oriented tangent line in each direction (see Figure 1b), and that has exactly one nonoriented tangent line in each direction if it is projective (see Figure 1c).



(a) (b) (c)

Figure 1

The results of this Note are based on the following lemma.

Lemma. *A projective hedgehog $\mathscr{H}_h$ has at least three singularities at the source $RP^1 = (S^1/antipodal)$.*

Proof: Since (3) implies

$$\frac{dx}{d\theta} = -(p + p'')(\theta)\sin\theta \text{ and } \frac{dy}{d\theta} = (p + p'')(\theta)\cos\theta,$$

the singularities of $\mathscr{H}_h$ correspond to the values of θ for which $(p + p'')(\theta) = 0$. Note that we may interpret the function $p + p''$ as the radius of curvature of $\mathscr{H}_h$.

Assume $p + p''$ has fewer than three zeroes on $[0, \pi]$. Since p is a Möbius function, so is $p + p''$. Thus $p + p''$ changes sign an odd number of times on $[0, \pi[$. Hence, it must have only one sign change on $[0, \pi[$, say at θ_0. Since the function $q(\theta) = \sin(\theta - \theta_0)$ changes sign only at θ_0 on $[0, \pi[$, $p + p''$ and q have on $[0, \pi[$ the same intervals of constant sign. Therefore

$$\int_0^\pi (p + p'')(\theta)q(\theta)\,d\theta \neq 0.$$

On the other hand, $q'' = -q$ and integrations by parts give

$$\int_0^\pi p''(\theta)q(\theta)\,d\theta = -\int_0^\pi p'(\theta)q'(\theta)\,d\theta = \int_0^\pi p(\theta)q''(\theta)\,d\theta.$$

Hence

$$\int_0^\pi (p + p'')(\theta)q(\theta)\,d\theta = 0,$$

a contradiction. ■

We know from the Four Vertex Theorem that a plane convex hedgehog has at least four vertices, that is, four points where the curvature has a stationary value. From the previous lemma, we can now deduce the following result.

Theorem 1. *A plane convex hedgehog of constant width has at least six vertices.*

Proof: The condition that a plane convex hedgehog is of constant width $2r$ (i.e., such that the distance between two parallel tangent lines is constant, equal to $2r$) is simply that its support function has the form $f + r$, where f is the support function of a projective hedgehog.

A vertex of $\mathscr{H}_{f+r}$ is a point of at least third order tangency of $\mathscr{H}_{f+r}$ with a circle, that is, a point where the radius of curvature of $\mathscr{H}_{f+r}$ has a stationary value. In other words, the vertices of $\mathscr{H}_{f+r}$ correspond to the values of θ for which $(p' + p''')(\theta) = 0$, where $p(\theta) = f(\cos\theta, \sin\theta)$. Since these values correspond to the singularities of the projective hedgehog defined by the support function $g(\cos\theta, \sin\theta) = p'(\theta)$, the theorem follows from the lemma. ■

Arnold's Tennis Ball Theorem **[1]** asserts that a closed simple smooth spherical curve dividing the sphere into two parts of equal areas has at least four inflection points. From our lemma, we can deduce the following result.

Theorem 2. *Let C be a closed simple smooth spherical curve that is everywhere transverse to the meridians and does not pass through the poles. If C is invariant under the antipodal map then C has at least six inflection points.*

Remark 1. *In these theorems, an inflection point is simply a zero of the geodesic curvature, that is, a point of at least second order tangency of the curve with a great circle.*

Remark 2. *We can also see Theorem 2 as a corollary of the classical Möbius theorem: a simple noncontractible curve in the projective plane has at least three inflection points.*

Proof of Theorem 2: Such a curve C has a parametrization of the form:

$$X_p: [0, 2\pi] \to S^2, \theta \mapsto \frac{(\cos\theta, \sin\theta, p(\theta))}{\sqrt{1 + p(\theta)^2}},$$

where p is a Möbius function.

If θ_0 is a zero of $p + p''$, we observe that the spherical curve parametrized by X_q, where $q(\theta) = p(\theta_0)\cos(\theta - \theta_0) + p'(\theta_0)\sin(\theta - \theta_0)$, is a great circle whose order of tangency with C at $X_p(\theta_0)$ is at least equal to two, since

$$q(\theta_0) = p(\theta_0), q'(\theta_0) = p'(\theta_0), q''(\theta_0) = -q(\theta_0) = -p(\theta_0) = p''(\theta_0).$$

Hence, the zeroes of $p + p''$ correspond to inflection points of C, and the theorem follows from the lemma. ∎

ACKNOWLEDGMENTS. I am most grateful to the referee who gave me the ideas to convert my first attempt into the Note in its final form. I also would like to thank H. Rosenberg for helpful comments.

REFERENCES

1. Arnold, V. I., *Topological Invariants of Plane Curves and Caustics*, AMS University lecture series, v. 5 (1994).
2. Langevin, R., Levitt, G., et H. Rosenberg, Hérissons et Multihérissons, (Enveloppes paramétrées par leur application de Gauss), Singularities, Warsaw (1985) 245–253. Banach Center Publ. 20, PWN Warsaw (1988).

1, rue Auguste Perret
92500 Rueil Malmaison
FRANCE

It is a person's taste in problems that decides what kind of mathematics he does....Peter Lax

D. Albers, G. Alexanderson, C. Reid (eds.),
***More Mathematical People*, Harcourt Brace Jovanovich,**
New York, 1990, p. 155.

UNSOLVED PROBLEMS

Edited by: **Richard Guy & Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

Bite-Sized Combinatorial Geometry Problems

Ben Goertzel, Computer Science, Waikato University, Hamilton, New Zealand, asks

> Are there three shapes, all different, so that any two of the shapes can be placed together to make the third shape? If so, what are the shapes? If not, why not?

A rigorous formulation of this problem is as follows. Let $I(A)$ denote the interior of a plane set A; define a **similitude** of the plane as an affine transformation that is a composition of a rotation, a translation, a magnification, and possibly a reflexion. Say that two sets are **similar** if one can be transformed into the other by a similitude. Are there three compact connected plane sets A, B, C so that

(1) A, B, C are mutually dissimilar, and
(2) There are similitudes L_i $(1 \leq i \leq 6)$ so that

$$A = L_1(B) \cup L_2(C), \qquad I(L_1(B)) \text{ and } I(L_2(C)) \text{ disjoint,}$$

$$B = L_3(C) \cup L_4(A), \qquad I(L_3(C)) \text{ and } I(L_4(A)) \text{ disjoint,}$$

$$C = L_5(A) \cup L_6(B), \qquad I(L_5(A)) \text{ and } I(L_6(B)) \text{ disjoint?}$$

If so, what are the shapes? If not, why not?

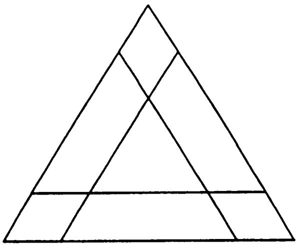
Note that, if we lift the restriction that the three shapes be different, then the puzzle has an easy solution: take three right-angled triangles of the same shape.

This **three shapes puzzle** naturally generalizes to an n-shapes puzzle: Are there n shapes, all different, so that any $n - 1$ of the shapes can be placed together to form the n-th shape?

James Gary Propp, Mathematics, Massachusetts Institute of Technology, proposes

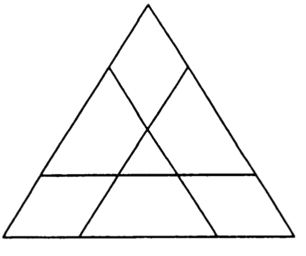
The Slicing Game

Here is a triangle that has been sliced three times:



Can you slice the triangle some more so that all the resulting pieces are triangles? All slice-lines must cut *completely* through the original triangle.

What if I make the starting configuration a little tighter, like this?



Or even tighter?

More generally, suppose that I have pre-sliced a convex polygon; can you always slice it further so that all the resulting pieces are triangles? (If the answer is "yes" then this would also apply to any "false starts" you may have tried and abandoned in attempting to solve the first two problems!)

Harry Tamvakis, Mathematics, The University of Chicago, originally asked if any convex polygon can be partitioned into a finite number of triangles of equal area and observed that the problem can be easily reduced to the following question: Is it possible to partition a convex quadrilateral of area 1 into a finite number of triangles of rational area?

[Not unrelated is Problem 37 in [2]: Let $ABCD$ be a convex quadrilateral. Find a necessary and sufficient condition for a point P to exist inside $ABCD$ such that the four triangles ABP, BCP, CDP, DAP all have the same area.

Answer. A necessary and sufficient condition is for one of the diagonals AC and BD to bisect the other.]

However, Tamvakis later observed that the negative answer to his question is given in [1]: there are quadrilaterals that cannot be partitioned into equal area triangles. Indeed, if we call these quadrilaterals **bad**, then Hales & Straus [3] and Kasimatis & Stein [5] show that the set of **good** quadrilaterals is small in the space of all quadrilaterals, both in the sense of measure (measure zero) and in the sense of category (first category). Their proofs are related to the work of Monsky [8] and involve Sperner's lemma and valuation theory. Can no one find elementary proofs? Kasimatis & Stein give a method of constructing examples of bad quadrilaterals: they show that if $C > 0$ is transcendental, then the trapezoid with vertices at $(0,0)$, $(1,0)$, $(0,1)$ and $(C,1)$ is bad: they ask what if $C = \sqrt{2}$?

Kasimatis [4] proves many intriguing theorems, including the following gem:

Theorem. *Let $n \geq 5$. A necessary and sufficient condition for a regular n-gon to have a partition into m equal area triangles is $n \mid m$.*

Note that this is false for $n = 4$.

Two other 'obvious' problems that have only recently been solved are confirmation of Klarner's 1969 conjecture that there is no polyomino of order three and that if a rectangle is dissected into three congruent pieces, the pieces must be rectangles. The **order** of a polyomino, which doesn't exist for most polyominoes, is the least number of copies that tile a rectangle. Ian Stewart [9] and his Patent Office friend [10] have settled Klarner's conjecture, and Sam Maltby [7] has generalized this result and his own result on dissecting a square [6] by proving the rectangle result. These proofs are all far from simple. Can they be simplified? Maltby asks: does his result hold for convex quadrilaterals? And can the result be generalized to higher dimensions?

REFERENCES

1. Croft, Hallard T., Falconer, Kenneth J. and Guy, Richard K., *Unsolved Problems in Geometry*, Springer, New York, 1991. **C6**. Cutting up squares, circles and polygons.
2. Gilbert, George T., Krusemeyer, Mark I. and Larson, Loren C., *The Wohascum County Problem Book*, Dolciani Math. Exposition **14** (1993) 10, 68–70, Math. Assoc. America, Washington, D.C.
3. Hales, A. W. and Straus, E. G., Projective colorings, *Pacific J. Math.* **99** (1982) 31–43; MR **83h**:51005.
4. Kasimatis, E. A., Dissections of regular polygons into triangles of equal areas, *Discrete Comput. Geom.* **4** (1989) 375–381.
5. Kasimatis E. A. and Stein, S. K., Equidissections of polygons, *Discrete Math.*, **85** (1990) 281–294.
6. Maltby, Samuel J., Solution to Problem 875, *Crux Mathematicorum* **17** (1991) 141–146.
7. Maltby, Samuel J., Trisecting a rectangle, *J. Combin. Theory Ser. A* **66** (1994) 40–52.
8. Monsky, Paul, On dividing a square into triangles, this MONTHLY **77** (1970) 161–164; MR **40** #5454.
9. Stewart, Ian N., *Another Fine Math You've Got Me Into*, W. H. Freeman, New York, 1992, Chapter 2, Tile and Error.
10. Stewart, Ian N. and Wormstein, Albert, Polyominoes of order 3 do not exist, *J. Combin. Theory Ser. A* **61** (1992) 130–136; MR **93d**:05039.

THE AUTHORS

RICHARD ASKEY took a course on differential equation from his coauthor's late husband when he was a freshman at Washington University. This was his first exposure to Bessel functions and some of the other functions he has studied for over 40 years. His first paper was a joint one with I. I. Hirschman, Jr., and dealt with ultraspherical polynomials. Both of these are relevant to the present paper.

DEBORAH TEPPER HAIMO completed her predoctoral work at Radcliffe College, which named her a distinguished alumna. A Harvard University Ph.D., she is now an Overseer there. Franklin and Marshall College gave her an honorary Dr. Sci. degree. A former MAA President, she is active in educational issues. As a classical analyst, she has published extensively in her field. She is a University of Missouri-St. Louis Professor Emerita; has held faculty posts at Southern Illinois University, Edwardsville, and at Washington University; and has been a member of the Institute for Advanced Study.

DICK WARREN wrote his Ph.D. thesis in general topology and then learned applied mathematics under the guidance of several scientists at Wright-Patterson AFB. This qualified him for technical positions in industry where he has led analyst teams that designed and prototyped new algorithms. His interest in the traveling salesman problem grew out of one of these efforts.

BENNETT EISENBERG attended Coolidge High School in Washington D.C., where he learned geometry from Carol McCammon. His interest in probability developed at Dartmouth College as an undergraduate research assistant to John Kemeny. After receiving his Ph.D. from M.I.T., he taught at Cornell University and the University of New Mexico before joining the mathematics department at Lehigh University in 1972. A problem from the 1992 Putnam exam sparked his interest in geometric probability.

ROSEMARY SULLIVAN received her bachelor's degree in mathematics from Penn State. She is currently pursuing a Ph.D. at Lehigh University. Probability problems with a geometric flavor are her main interest. When not doing mathematics, she enjoys reading, hiking, running, and skiing.

ADRIAN OLDKNOW is Professor of Mathematics and Computing Education at the Chichester Institute of Higher Education, UK. He has degrees in Mathematics from Oxford University, and in Computer Science from Brunel University. He has taught in secondary schools, tertiary colleges, teacher-training colleges, and the university sector. One of his main concerns is to review the way that computers, and sophisticated software, are changing the practice of mathematics in industry, commerce, and research–and the implications for the mathematical education of teachers, undergraduates, and school students. He has had a life-long passion for geometry (both of the old-fashion and the CAD type).

SAUNDERS MAC LANE Ph.B. Yale 1930, Ph.D. Göttingen 1934; President, MAA 1950–51; co-author with Garrett Birkhoff "A Survey of Modern Algebra" (Macmillan, 1941). Taught at Harvard, Cornell, and the University of Chicago. Now Professor Emeritus, Chicago.

URI LERON completed his Ph.D. in 1972 under the supervision of S. A. Amitsur at the Hebrew University of Jerusalem. The next ten years he taught and did research in ring theory at the University of Oregon, UCLA, and the Israel Institute of Technology. He has held visiting positions at MIT and U. C. Berkeley. Since 1980 he became seriously involved in research and development in math and computer science education.

ED DUBINSKY spent 23 years in functional analysis research. Since 1985 he has done research and development in undergraduate math education. He investigates how to help students make the mental constructions that can lead to understanding and using math concepts. He has founded and (co-)directed institutes in computer science education, calculus, and high school math. He is editor of *UME Trends*, chair of the Joint Committee on Research in Undergraduate Mathematics Education, and co-editor of *Research in Collegiate Mathematics Education.*

PETER HILTON was born in London, England. He had a career in the U. K. (including service in WWII breaking German codes in a team that included Alan Turing) and came to the United States in 1962. His principal research interests are algebraic topology, homological algebra, and group theory (in which fields he has published 7 books and some 400 articles), but he also has a strong concern to try to improve the quality of mathematical education at all levels. He is currently Distinguished Professor Emeritus at the State University of New York (Binghamton) and Distinguished Service Professor at the University of Central Florida (Orlando).

We may safely say that the whole form of modern mathematical thinking was created by Euler. It is only with the greatest difficulty that one is able to follow the writings of any author immediately preceding Euler, because it was not yet known how to let the formulas speak for themselves. This art Euler was the first one to teach.

R. Rubio

Euler could repeat the Aeneid from the beginning to the end, and he could even tell the first and last lines in every page of the edition which he used. In one of his works there is a learned memoir on a question of mechanics, of which, as he himself informs us, a verse of the Aneid ["The anchor drops, the rushing keel is staid."] gave him the first idea.

David Brewster

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS *address given on the inside front cover. Please include solutions and relevant references. Two copies of all items needed to evaluate the problem should be sent. A third copy of the the problem and solution is often useful; please include one if possible.*

Solutions of published problems should arrive at the MONTHLY PROBLEMS *address given on the inside front cover before September 30, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.*

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

A typographical error in problem 10491 [1995, 930] led to the publication of a significantly different statement from what had been proposed. The correct statement is given below, before the regularly scheduled problems. The deadline for solutions to 10491 is extended to September 30, 1996.

10491. *Proposed by Jean-Pierre Grivaux, Lycée Chaptal, Paris, France.*

Let B be an open ball containing the origin in the Euclidean space $\mathbb{R}^n$, and let V denote its volume. B is cut into 2^n parts by the coordinate hyperplanes

$$\Pi_i = \{(x_1, \ldots, x_n) : x_i = 0\}$$

for $i = 1, \ldots, n$. Prove that at least 2^{n-1} of these parts have volume at most $V/2^n$.

10515. *Proposed by Albert Wilansky, Lehigh University, Bethlehem, PA.*

For which values of m and n (each at least 2) will a *reflecting bishop* placed on a white corner square of an otherwise empty m by n chessboard attack every white square on the board (except for the one it occupies).

10516. *Proposed by Donald A. Darling, Newport Beach, CA.*

Let (X, Y, Z) be three random variables such that $\alpha X + \beta Y + \gamma Z$ is uniformly distributed over the interval $[-1, 1]$ for every set of three *direction cosines*, i.e., numbers with $\alpha^2 + \beta^2 + \gamma^2 = 1$. Show that $X^2 + Y^2 + Z^2 = 1$ with probability one.

10517. *Proposed by Jean Anglesio, Garches, France.*

Let ΔABC be a triangle and let H be its orthocenter and I its incenter. If W is the point such that $\overrightarrow{HW} = 4\overrightarrow{HI}$ and $R = 2\,|HI|\,\sqrt{2}$, show that none of the vertices A, B or C is in the interior of the circle with center W and radius R.

10518. *Proposed by Yuanan Diao, Kennesaw State College, Marietta, GA.*

Let X be a finite set of points in a metric space and let X_1 and X_2 be a partition of X into two disjoint nonempty sets. Let

$$d(X_1, X_2) = \min\{d(x_1, x_2) : x_1 \in X_1, x_2 \in X_2\}$$

be called the distance between the subsets, and let the largest value of the distance between two such subsets be called the *splitting number* of X.

If X consists of n random points, independently selected from the uniform distribution on a ball of radius 1 in 3-dimensional Euclidean space, show that the splitting number of X is almost surely small. More precisely, for $a < 1$, show that there is a constant $\alpha > 0$ depending only on a that the splitting number of X is less than a with probability at least $1 - e^{-\alpha n}$.

10519. *Proposed by Jürgen Groß, Götz Trenkler, and Sven-Oliver Troschke, University of Dortmund, Dortmund, Germany.*

Let A be an n by n complex matrix. Denote its conjugate transpose by A^*. If $A^2 = A$, show that $A = A^*$ if and only if Range A = Range A^*.

10520. *Proposed by Clark Kimberling, University of Evansville, Evansville, IN.*

Suppose α is a real irrational number greater than 1. For all i and j in the set $\mathbb{N}$ of positive integers, let

$$a(i, j) = \sum_{k=1}^{\infty} \left\lfloor i\alpha^{j-k} \right\rfloor.$$

(a) Prove that a is a one-to-one correspondence from $\mathbb{N} \times \mathbb{N}$ onto $\mathbb{N}$.

(b) For $n = a(i, j) \in \mathbb{N}$, let $s_n = i$. Exhibit the sequence $\langle s \rangle$ as a proper subsequence of itself.

10521. *Proposed by D. M. Bloom & G. W. Booth, Brooklyn College, CUNY, Brooklyn, NY.*

Let $S_n = (2\pi n)^{1/2}(n/e)^n$ and $T_n = n!/S_n$.

(a) Prove that

$$T_n - 1 = \frac{1}{12n - a_n}$$

where $0 < a_n < 1/2$ for all positive integers n.

(b) Prove that the sequence $\langle a \rangle$ is monotonically increasing.

(c)* If $b_n = n(1/2 - a_n)$ for all $n \in \mathbb{N}$, is the sequence $\langle b \rangle$ monotonically increasing?

NOTES

(**10515**) The *reflecting bishop* has appeared in *Fairy Chess* problems. When it reaches a (non-corner) edge square, it may "reflect" onto the other diagonal passing through that square. (**10517**) The triangle should be assumed to be nondegenerate, i.e., A, B, and C should not be collinear. (**10519**) The matrix A is said to be *range-Hermitian* if Range A = Range A^*. Thus, the the problem asserts that *A is a Hermitian projection if and only if A is a range-Hermitian projection.* (**10520**) The explicit self-similarity of the sequence $\langle s \rangle$ suggests that it could be considered to be a *fractal sequence.* (**10521**) One recognizes S_n as the Stirling approximation to $n!$, so it is known that $\lim_{n\to\infty} T_n = 1$. A weaker form of (a), with $-1 < a_n < 1$, appears in Taylor and Mann, *Advanced Calculus*, 3rd ed., p. 705. In addition, it follows from the asymptotic expansion of the logarithm of the Gamma function (see E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Fourth Edition, Cambridge, 1927, sect. 12.33) that $\lim_{n\to\infty} a_n = 1/2$ and $\lim_{n\to\infty} b_n = \frac{293}{720}$. However, the questions here concern the behavior of the elements of the sequence rather than only the limits.

SOLUTIONS

The Number of Almost-linear Sequences

10302 [1993, 401]. *Proposed by Jeffrey A. Barnett, Northrop Corporation, Palos Verdes Peninsula, CA.*

Let $\mathfrak{J}$ be the class of integer-valued functions defined on $\mathbb{N}^+$ that satisfy the following constraints. For $j \in \mathfrak{J}$, $j(1) = 1$ and

$$j(m) + j(n) - 1 \le j(m+n) \le j(m) + j(n)$$

for all $m, n \in \mathbb{N}^+$. Show that the number of distinct initial sequences of length N generated by the $j \in \mathfrak{J}$ is

$$\sum_{1\le n\le N} \phi(n)$$

where $\phi(n)$ is Euler's totient function.

Solution by Robin J. Chapman, University of Exeter, Exeter, U. K.. For $0 < x \le 1$, define j_x by $j_x(n) = \lceil nx \rceil$. Clearly $j_x(1) = 1$ and $j_x(n) \in \mathbb{N}^+$ for all $n \in \mathbb{N}^+$. Since $j_x(n) = nx + a_n$ for some $a_n \in [0, 1)$, we have $j_x(m) + j_x(n) - j_x(m+n) = a_m + a_n - a_{m+n} \in (-1, 2)$. Since this is an integer, it must be 0 or 1, and we have $j_x \in \mathfrak{J}$.

The set A_N of rational numbers in $(0, 1]$ expressible as fractions with denominators at most N has $\sum_{1\le n\le N} \phi(n)$ elements. We establish a bijection between A_N and the set of initial sequences of length N generated by the sequences in $\mathfrak{J}$.

Suppose x, y are distinct elements of A_N with $x < y$. If $x \in A_N$ has denominator $n \le N$, then $j_x(n) = nx$, but $j_y(n) \ge ny > nx$. Hence $j_y(n) \ne j_x(n)$, and j_x, j_y have distinct initial sequences of length N.

Given $j \in \mathfrak{J}$, we next obtain $x \in A_N$ such that $j(n) = j_x(n)$ for $1 \le n \le N$. Let $x = \min_{1\le n\le M} j(n)/n$, and choose $m \le N$ with $x = j(m)/m$. Clearly $x \in A_N$. By the choice of x, for $1 \le n \le N$ the number $j(n)$ is an integer at least nx; hence

$j(n) \geq \lceil nx \rceil = j_x(n)$. To prove j agrees with j_x through N, suppose there exists $n \leq N$ such that $j(n) > j_x(n) = \lceil nx \rceil$, so $j(n) \geq nx + 1$. When we express mn as n copies of m, the defining condition yields $j(mn) \leq nj(m) = nmx$. When we express mn as m copies of n, the defining condition yields $j(mn) \geq mj(n) - (m-1) \geq mnx + 1$, a contradiction.

Editorial comment. Frank Schmidt pointed out that these sequences are characterized in R. L. Graham, S. Lin, and C.-S. Lin, "Spectra of numbers", *Math. Mag.* 51 (1978), 174–176. Another reference, supplied by an editor, is M. Boshernitzan and A. S. Fraenkel, "Nonhomogeneous spectra of numbers", *Discr. Math.* 34 (1981), 325–327.

Solved also by V. Božin (student, Yugoslavia), H. von Eitzen (Germany), F. J. Flanigan, R. Holzsager, I. Kastanas, O. P. Lossers (The Netherlands), A. D. Melas (Greece), J. M. Santmyer, F. Schmidt, A. N. 't Woord (The Netherlands), GCHQ Problem Solving Group (U. K.), and the proposer.

A Modular Power Series

10314 [1993, 589]. *Proposed by Andrew Vince, University of Florida, Gainesville, FL.*

Let b be an integer greater than 1. Let S be a set of integers containing 0 such that no two members of S are congruent modulo b. If

$$\sum_{i=1}^{\infty} \frac{s_i}{b^i} = 0,$$

with $s_i \in S$, prove that all $s_i = 0$.

Solution by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. Suppose that $\sum_{i=1}^{\infty}(s_i/b^i) = 0$ with $s_i \in S$. We define a sequence $\{a_n\}_{n=0}^{\infty}$ such that $\sum_{i=1}^{n}(s_i/b^i) = a_n/b^n$, by setting $a_0 = 0$ and $a_n = b^{n-1}s_1 + b^{n-2}s_2 + \cdots + bs_{n-1} + s_n$ for $n \geq 1$. Notice that $a_n = ba_{n-1} + s_n$ for all $n \geq 1$.

If $a_n = 0$ for some $n \geq 1$, then $s_n \equiv a_n \equiv 0 \bmod b$. This requires $s_n = 0$ and $a_{n-1} = 0$, and hence $s_i = 0$ for all $i \leq n$, inductively. If $1 \leq m < n$ and $a_m = a_n$, then $s_m \equiv a_m \equiv a_n \equiv s_n \bmod b$. Therefore, $s_m = s_n$ and $a_{m-1} = a_{n-1}$. By induction, $a_{n-m} = a_0 = 0$. Hence $s_i = 0$ for all $i \leq n - m$. It suffices therefore to show that $a_m = a_n$ occurs with the difference $n - m$ arbitrarily large.

By the congruence condition, S is finite. Choose $M > 0$ such that $|s_i| \leq M$ for all i. Now

$$0 = \left|\sum_{i=1}^{\infty} \frac{s_i}{b^i}\right| \geq \frac{|a_n|}{b^n} - \sum_{i=n+1}^{\infty} \frac{M}{b^i} = \frac{1}{b^n}\left(|a_n| - \frac{M}{b-1}\right),$$

so $|a_n| \leq M/(b-1)$ for all n. Therefore, the set $\{a_n : n \geq 0\}$ is finite. Thus, for some k, we have $a_n = k$ infinitely often, and the result follows.

Solved also by K. L. Bernstein, W. Blumberg, S. M. Gagola Jr., R. Holzsager, N. Jensen (Germany), I. Kastanas, O. P. Lossers (The Netherlands), R. Martin (student), L. E. Mattics, the MMRS group of Oklahoma State University, and the proposer. Five incorrect solutions were received.

Integral Matrices with Integral Inverses

10315 [1993, 589]. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Let A and B be matrices with integer entries of sizes r by n and n by r, respectively, with $r < n$. Suppose that AB is an r by r identity matrix. Show that A can be enlarged to an n by n integral matrix having an integral inverse.

Solution by Allan Pedersen, Søborg, Denmark. Let $a_1, \ldots, a_r$ be the row vectors of A in order from top to bottom, and let $b_1, \ldots, b_r$ be the column vectors of B in order from left to

right. It is given that $a_i \cdot b_j = \delta_{ij}$ for $1 \le i, j \le r$. It suffices to obtain vectors a_{r+1}, b_{r+1} such that $a_i \cdot b_j = \delta_{ij}$ for $1 \le i, j \le r+1$.

Since $r < n$, the system of linear equations $x \cdot b_j = 0$ for $1 \le j \le r$ has a nontrivial rational solution. Let $a_{r+1} \neq 0$ be such a solution. We may assume that the entries of a_{r+1} have greatest common divisor 1. Hence there is a vector b such that $a_{r+1} \cdot b = 1$. Let $b_{r+1} = b - \sum_{1=j}^{r}(a_j \cdot b)b_j$. Direct computation verifies that a_{r+1}, b_{r+1} provides the desired enlargement.

Editorial comment. Readers found the result in various places in the literature. Allan Pedersen provided the oldest reference: F. G. Frobenius, "Theorie der linearen Formen mit ganzen Coefficienten", *J. Reine Angew. Math.* 86(1879), 146–208, (reprinted in Ferdinand Georg Frobenius, *Gesammelte Abhandlungen* (J.-P. Serre, ed.), Springer, 1968). He described the above proof as essentially that of Frobenius. The other proofs used a large variety of results in linear algebra.

Solved also by R. J. Chapman (U. K.), T. H. Foregger, F. Gaines (Ireland), R. Holzsager, N. Jensen (Germany), I. Kastanas, C. Lanski, J. G. Mauldon, W. A. McWorter, Jr. & L. F. Meyers, J. F. Queiró (Portugal), R. B. Richter (Canada) & W. P. Wardlaw, F. Schmidt, R. B. Tucker, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, the MMRS group of Oklahoma State University, National Security Agency Problems Group, and the proposer. One incorrect solution was received.

Positive Quadrant Dependent Random Variables

10325 [1993, 689]. *Proposed by Broderick Oluyede, Georgia State University, Atlanta, GA.*

For $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, let $p_{i,j} \ge 0$, and assume that

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{i,j} = 1.$$

Define $p_{i,\cdot} = \sum_{j=1}^{c} p_{i,j}$ and $p_{\cdot,j} = \sum_{i=1}^{r} p_{i,j}$. In addition, suppose that

$$p_{i+1,j+1}\sum_{h=1}^{i}\sum_{k=1}^{j} p_{h,k} \ge \sum_{h=1}^{i} p_{h,j+1}\sum_{k=1}^{j} p_{i+1,k}$$

for $0 < i < r$ and $0 < j < c$. Prove that

$$\sum_{h=1}^{i}\sum_{k=1}^{j} p_{h,k} \ge \sum_{h=1}^{i} p_{h,\cdot}\sum_{k=1}^{j} p_{\cdot,k}$$

for $0 < i < r$ and $0 < j < c$.

Solution by Richard Holzsager, The American University, Washington, DC. Write $S_{i,j}$ for the sum $\sum_{h=1}^{i}\sum_{k=1}^{j} p_{h,k}$. Then the hypothesis is easily seen to be equivalent to $S_{i,j}\, S_{i+1,j+1} \ge S_{i,j+1}\, S_{i+1,j}$ for $0 < i < r$ and $0 < j < c$ and the conclusion to $S_{i,j}\, S_{r,c} \ge S_{i,c}\, S_{r,j}$ for $0 < i < r$ and $0 < j < c$. We generalize to $S_{i,j}\, S_{m,n} \ge S_{i,n}\, S_{m,j}$, $0 < i < m \le r, 0 < j < n \le c$.

Since the $p_{h,k} \ge 0$, the only way that one could have $S_{i,j} = 0$ is for all $p_{h,k} = 0$ for $0 < h \le i$, $0 < k \le j$. Collecting together these zeros gives $S_{h,k} = 0$. Furthermore, $S_{i,j} = 0$ implies that $S_{i,j+1}\, S_{i+1,j} = 0$, so that either $S_{i,j+1} = 0$ or $S_{i+1,j} = 0$. Continuing in this way, one finds that either $S_{i,n} = 0$ or $S_{m,j} = 0$, so the result is true if $S_{i,j} = 0$. We now assume $S_{i,j} \neq 0$ and prove the result by induction on $m - i$ and $n - j$.

The case $m - i = n - j = 1$ was given. Assume, say, that $m - i \ge 2$ and that, by induction, $S_{i+1,j}\, S_{m,n} \ge S_{i+1,n}\, S_{m,j}$ and $S_{i,j}\, S_{i+1,n} \ge S_{i,n}\, S_{i+1,j}$. (A similar construction will apply when $n - j \ge 2$). Multiply the first inequality by $S_{i,j}$ and the second by $S_{m,j}$, and

add to get $S_{i,j}\,S_{i+1,j}\,S_{m,n} + S_{i,j}\,S_{i+1,n}\,S_{m,j} \geq S_{i,j}\,S_{i+1,n}\,S_{m,j} + S_{i,n}\,S_{i+1,j}\,S_{m,j}$. This simplifies to $S_{i,j}\,S_{i+1,j}\,S_{m,n} \geq S_{i,n}\,S_{i+1,j}\,S_{m,j}$. Since $S_{i+1,j} \neq 0$, we can cancel this factor and complete our inductive step.

Editorial comment. All solvers noted that the expression "$0 < j < r$" in the original statement was a misprint. This has been corrected above. Although it plays no role in the solution, the $p_{i,j}$ may be considered as probabilities. David Aldous noted that the property proved of these probabilities has been studied under the name given as a title to this problem, and he supplied references to Richard E. Barlow & Frank Proshan, *Statistical Theory of Reliability and Life Testing Probability Models*, Holt, Rinehart & Winston, 1975 and Henry W. Block, Allan R. Sampson & Thomas H. Savits (eds.), Topics in Statistical Dependence, Inst. Math. Statist., 1991.

Solved also by O. P. Lossers (The Netherlands), M. Shemesh (Israel), A. N. 't Woord (The Netherlands), and the proposer.

Euler's Theorem Implies the Pythagorean Proposition

Let a, b, c be the sides of a right triangle, placed on the x-axis, with θ, the angle between the sides a and c, placed at the origin, and the side a on the x-axis. Then Euler's Formula gives:

(1) $$a + bi = ce^{i\theta}.$$

Taking the complex conjugate, we get:

(2) $$a - bi = ce^{-i\theta}.$$

Multiplying (1) and (2) yields

(3) $$a^2 + b^2 = c^2.$$

Contributed by Maurice Machover, St. John's University, Jamaica, NY.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

A Mathematician Reads the Newspaper. By John Allen Paulos. Basic Books, 1995, vi + 200, $18.00

Reviewed by **Peter Hilton**

The author, John Allen Paulos, is Professor of Mathematics at Temple University in Philadelphia. His research is in the areas of probability theory and logic. He is also a highly effective popularizer of mathematics, laying stress on its importance and on the disadvantages experienced by the individual who is unable to reason mathematically. Two of his books, *Innumeracy* and *Beyond Numeracy*, were highly successful and deservedly so.

In this book, his avowed purpose is to show how mathematical literacy can enable us to read the typical American newspapers with greater insight and awareness of the risks of being misinformed. Generally speaking, he is really discussing the capacity for quantitative reasoning rather than mathematical aptitude, although logical considerations are also applied[1] and, in one key section—of which more anon—he introduces the reader, very appropriately, to chaos theory. Naturally, the misuse, deliberate or inadvertent, of statistics is his principal target.

He organizes the book into 5 sections, each consisting of about ten articles. The articles are very short—some only one page in length and none more than seven pages—and would appear at first to be collected from articles published in high-grade journals; this impression, however, is presumably false. Each article has a title and a subtitle, and discusses some type of newspaper article and its potential to mislead. The **title** is a typical newspaper headline introducing such newspaper articles and the **subtitle** characterizes the content and nature of such newspaper articles. Thus—and this was the first source of difficulty for the reviewer—there may well be no reference whatever in the course of an article in this book to the items featuring in its title. Thus the article 'Pakistan's Bhutto Gambles in Trade Negotiations' does not refer to Benazir Bhutto; and the article 'Near-Perfect Game for Roger Clemens' does not refer to Roger Clemens. Moreover, this labeling technique leads to the possibility of titles of highly ephemeral significance (dare one suggest that the two examples above may exhibit this danger?), which would have the effect of making it difficult for the future reader to anticipate the thrust of the article.

Most of the articles are very good. Some are immensely important, such as the one discussing the popular aversion to numbers ('Iraqi Death Toll Unknown') and

[1]The logic is virtually all about self-referential statements. There are 8 references to 'self-reference' in the Index (p. 211). The reviewer would have added a ninth—p. 211.

that discussing health hazards ('Ranking Health Risks: Experts and Laymen Differ'); it is no coincidence that the latter is the longest in the book.

For the brevity of the articles is, in the reviewer's judgment, a serious barrier to their effectiveness. It is not easy for the reader to grasp the various points being made—and the points made are various and by no means confined to those to be expected of a critical mathematician—when they are made so succinctly. In any presentation there is, of course, a trade-off between brevity and intelligibility; dare one suggest that the author here is more certain to be popular than to be understood?

The problem of intelligibility is further complicated by the author's indulgence of his considerable wit. There are many anecdotes and diverting asides—but some of these come close to being irrelevant, and this is a serious matter in such concise presentations. The long charming footnote on p. 86 seems not to relate at all to the theme of the article; moreover, Pepsi's catastrophic lottery, occupying a 13-line paragraph on the second page of a 3-page article is not an example of the subtitle 'Advertising and Numerical Craftiness'. Again, on p. 102, the author discusses, very relevantly, the celebrity's paradox 'I am famous', but precedes it by remarking on the logical difficulty of the self-referential statement 'I am lying'.

The quest for brevity and pithiness has also led the author, in some places, to fall below his own high standard of exposition. He breaks the Golden Rule of Pedagogy by asking the reader, in some places, to accept his authority without being offered any explanation. The most glaring example is to be found in the article beginning on p. 160, subtitled 'Ecology, Chaos and the News'. This has the makings of a brilliant exposition of a matter of great importance—that chaos can arise without human intervention—but the argument is the most mathematical of the entire book and it is not given in anything like sufficient detail to be intelligible to the bright layman. In 3 pages,[2] he or she is asked to understand the behaviour of the logistic function $f(x) = Rx(1 - x)$ on the open unit interval for values of R between 0 and 4. Why choose these values of R? (Because then f maps the unit interval to itself and so can be iterated.) If $R = 1.5$, why do the iterates stabilize at $\frac{1}{3}$? (Draw a picture, and note that $1 - 1/R$ is a fixpoint which lies in the interval provided $R > 1$.) If $R = 2.5$, why do the iterates stabilize at $\frac{3}{5}$? (Similarly.) If $R = 3.2$, why don't the iterates stabilize at the fixpoint $\frac{11}{16}$? (Draw a picture and discuss the difference between attractors and repellers.) How can any of the far more chaotic behaviour as R approaches 4 be understood if nothing of this is even mentioned? To popularize mathematics it is neither necessary nor sufficient to suppress it.

But no review of this worthy effort should end on a sour note. This book will bring a great deal of pleasure to many—as it did to the reviewer. It is full of fun, full of information, full of insights. It gives a thoroughly convincing answer to those who may question the importance of quantitative thinking and, indeed, of mathematics itself. For doing just this the author would deserve our warm thanks.

Department of Mathematics
University of New England
Armidale, NSW 2351
AUSTRALIA

[2]But half a page is taken up with a somewhat inscrutable diagram.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, L*, P*. *Lion Hunting & Other Mathematical Pursuits.* Eds: Gerald L. Alexanderson, Dale H. Mugler. Dolciani Math. Expos., V. 15. MAA, 1995, xii + 308 pp, $35 (P). [ISBN 0-88385-323-X] From the Introduction: "... an assortment of many of his [Ralph P. Boas, Jr.] lighter mathematical papers, along with verse, stories, anecdotes, and recollections." Includes mathematical articles (on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, and polynomials), articles about the teaching of mathematics, and a bibliography of Boas' mathematical books and articles. AO

General, L.** *She Does Math! Real-Life Problems from Women on the Job.* Ed: Marla Parker. MAA, 1995, xv + 253 pp, $24 (P). [ISBN 0-88385-702-7] Career histories of 38 professional women and math problems written by them. Intended to motivate students to take math throughout high school, and to encourage them to consider careers in technical fields. AO

General, P, L. *The Encyclopedia of Integer Sequences.* N.J.A. Sloane, Simon Plouffe. Academic Pr, 1995, xiii + 587 pp, $44.95. [ISBN 0-12-558630-2] Expanded version of *A Handbook of Integer Sequences* (TR, April 1973). Contains 5488 sequences, each with a brief description and a reference. A valuable resource. CEC

General, P*, L. *A Practical Guide to Cooperative Learning in Collegiate Mathematics.* Eds: Barbara E. Reynolds, *et al.* MAA Notes No. 37. MAA, 1995, xii + 176 pp, $18.95 (P). [ISBN 0-88385-095-8] "Nuts and bolts" advice on using cooperative learning. Presents techniques for forming groups, examples of group tasks, assessment strategies, suggestions about group dynamics, and information about student reactions. Also includes brief descriptions of other cooperative learning models and an extensive, annotated bibliography. AO

Reference, L, P. *The Advanced TEXbook.* David Salomon. Springer-Verlag, 1995, xx + 490 pp, $39.95 (P). [ISBN 0-387-94556-3] For experienced users who want to exploit more of TEX's capabilities. Covers some of TEX's advanced features including macros, conditionals, file I/O, leaders, the line- and page-break algorithms, and output routines. Numerous examples and exercises (most with solutions). AO

Reference, L*, P.** *Math into LATEX: An Introduction to LATEX and AMS-LATEX.* George Grätzer. Birkhäuser Boston, 1996, xxvii + 451 pp, $49.50 (P). [ISBN 0-8176-3805-9] First part is a short course for beginners; rest is an easy-to-read, systematic introduction to major features of the current versions of LATEX and AMS-LATEX. (An appendix discusses converting from older versions.) Numerous examples. Based on the author's earlier book *Math into TEX: An Introduction to AMS-LATEX* (TR, October 1994). AO

Reference, S(13–15), P. *Doing Mathematics with Scientific WorkPlace.* Darel W. Hardy, Carol L. Walker. Brooks/Cole, 1995, xix + 340 pp, $25.95 (P). [ISBN 0-534-34049-0] Nice tutorial overview of *Scientific WorkPlace* (a scientific text processing package integrated

with *Maple* and LaTeX). Oriented towards undergraduate student use. RM

Mathematics Appreciation, T(14–16: 1).** *The Mathematical Experience, Study Edition.* Philip J. Davis, Reuben Hersh, Elena Anne Marchisotto. Birkhäuser Boston, 1995, xxiii + 487 pp, $38.50. [ISBN 0-8176-3739-7]; *The Companion Guide, Study Edition*, vi + 120 pp, $14.95 (P). [ISBN 0-8176-3849-0] Original 1980 text (TR, May 1981; Extended Review, August–September 1982) augmented by exercises, essay topics, and bibliographies. *Companion Guide* contains discussion guides, topics for expository research papers, projects, sample exams, etc. AO

Mathematics Appreciation, T(13–14: 1). *Excursions in Modern Mathematics, Second Edition.* Peter Tannenbaum, Robert Arnold. Prentice Hall, 1995, xv + 598 pp. [ISBN 0-13-386921-0] New edition keeps material current, is more readable (*First Edition*, TR, May 1992). More, and new types of examples. Exercises refined, expanded. Student supplement contains recent articles from *New York Times*. JCS

Education, P. *Complex Problem Solving: The European Perspective.* Eds: Peter A. Frensch, Joachim Funke. Lawrence Erlbaum Assoc, 1995, xv + 340 pp, $34.50 (P); $59.95. [ISBN 0-8058-1783-2; 0-8058-1336-5] Novelty, complexity, dynamic change, and "intransparency" distinguish complex (CPS) from simple (SPS) problem solving. North American research, pioneered by Herbert Simon, has focused on domain-specific approaches (e.g., chess, law, writing); European research—the subject of this anthology—takes a more global approach, often relying on sophisticated computer models. Both seek to understand CPS and how it develops, since it is not just the natural consequence of extensive SPS. LAS

Education, P. *The Emergence of Mathematical Meaning: Interaction in Classroom Cultures.* Ed: Paul Cobb. Stud. in Math. Thinking & Learning Ser. Lawrence Erlbaum Assoc, 1995, xi + 306 pp, $27.50 (P); $59.95. [ISBN 0-8058-1729-8] Essays on various aspects of constructivism techniques in early education from groups in the United States and Germany. PF

History, P. *The Collected Papers of Albert Einstein, Volume 5: The Swiss Years Correspondence, 1902–1914.* Transl: Anna Beck. Consultant: Don Howard. Princeton Univ Pr, 1995, xxii + 384 pp, $29.95 (P). [ISBN 0-691-00099-9] New English translation of Einstein correspondence. Unfortunately, it does not contain annotations included in the original version. Interesting exchanges with Planck, Sommerfeld, and, especially, Lorentz. SK

Logic, T(13–15: 1, 2), S. *Introduction to Logic and to the Methodology of Deductive Sciences.* Alfred Tarski. Transl: Olaf Helmer. Dover, 1995, xvi + 239 pp, $8.95 (P). [ISBN 0-486-28462-X] Unabridged reprint of the 1946 *Revised Second Edition*, published by Oxford Univ. Press. Part 1 treats the elements of logic and the deductive method; Part 2 addresses applications of logic and methodology for constructing mathematical theories. RJA

Group Theory, T(15–16: 1), S, P, L. *Algebra and Tiling: Homomorphisms in the Service of Geometry.* Sherman K. Stein, Sándor Szabó. Carus Math. Mono., No. 25. MAA, 1994, xii + 207 pp, $34. [ISBN 0-88385-028-1] A lovely book demonstrating how group theory can be used to solve subtle problems in geometry. BC

Algebra, S(17–18), P. *Skew Fields: Theory of General Division Rings.* P.M. Cohn. Ency. of Math. & Its Applic., V. 57. Cambridge University Pr, 1995, xv + 500 pp, $89.95. [ISBN 0-521-43217-0] Comprehensive, definitive treatment, extends and updates *Skew Field Constructions* (TR, November 1977): free ideal rings (firs), localization, Bergman's coproduct construction, relations, equations, singularities, and valuations. RM

Calculus, T(13). *Brief Calculus: Applications + Technology.* Edmond C. Tomastik. Saunders College, 1996, xxiii + 554 pp, $27 (P). [ISBN 0-03-006868-1] Standard brief calculus text incorporating graphing calculators throughout. Worth considering. PF

Complex Analysis, T(17–18: 1), S, P, L. *Complex Dynamics and Renormalization.* Curtis T. McMullen. Annals of Math. Stud., No. 135. Princeton Univ Pr, 1994, vii + 214 pp, $22.50 (P); $49.50. [ISBN 0-691-02981-4; 0-691-02982-2] An excellent account of the mathematical methods that have been developed to give rigorous proofs of properties of the Mandelbrot set, Julia sets, and other "complex" geometric objects. BC

Differential Equations, T(17: 1, 2). *Delay Equations: Functional-, Complex-, and Nonlinear Analysis.* Odo Diekmann, *et al.* Appl. Math. Sci., V. 110. Springer-Verlag, 1995, xi + 534 pp, $49. [ISBN 0-387-94416-8] Qualitative theory of delay differential equations—contains all the operator theory a differential equator needs to understand infinite dimensional state spaces, and all the dynamical systems theory an analyst would need. SK

Differential Equations, T(14: 1). *Differential Equations: Matrices and Models.* Paul Bugl. Prentice Hall, 1995, xvi + 669 pp. [ISBN 0-02-316540-5] From the Preface: "The ultimate goal of this book is the use of matrix methods for the solution of systems of linear ordinary differential equations." Some traditional topics are omitted or given short shrift (e.g., exact equations, undetermined coefficients); others are given considerable attention (e.g., series solutions). Includes some not-so-traditional topics (e.g., ill-conditioned systems, boundary and initial value Green's functions). Models from classical mechanics and electrical circuit theory. AO

Differential Equations, T*(14: 1). *Modern Differential Equations: Theory, Applications, Technology.* Martha L. Abell, James P. Braselton. Saunders College, 1996, xviii + 627 pp. [ISBN 0-03-098337-1] Traditional content presented in a way that takes advantage of modern technology. Examples, exercises, and projects based on applications in a wide variety of fields including biology, chemistry, physics, economics, and engineering. AO

Dynamical Systems, T(16–17: 1, 2), S, P, L. *Applied Nonlinear Dynamics: Analytical, Computational, and Experimental Methods.* Ali H. Nayfeh, Balakumar Balachandran. Ser. in Nonlinear Sci. Wiley, 1995, xv + 685 pp, $64.95. [ISBN 0-471-59348-6] Treats equilibrium, periodic, and quasiperiodic solutions of systems of ODE's, as well as chaos. Concludes with a chapter on control of bifurcations, chaos control, and synchronization. BC

Dynamical Systems, P. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems.* Hal L. Smith. Math. Surveys & Mono., V. 41. AMS, 1995, x + 174 pp, $49. [ISBN 0-8218-0393-X] A monotone dynamical system is a partially ordered metric space and a semi-flow on it that preserves the partial ordering. A system of ODE's is cooperative (resp., competitive) if the forward (resp., backward) time semi-flow is monotone. Monotonicity allows one to ratchet down a dimension, so that, for example, one can apply the Poincaré–Bendixson theorem to compact limit sets of three-dimensional competitive and cooperative systems. Applications also to delay equations. SK

Numerical Analysis, T(15–17:1). *Numerical Linear Algebra and Applications.* Biswa Nath Datta. Brooks/Cole, 1995, xxii + 680 pp. [ISBN 0-534-17466-3]

Analysis, P. *Harmonic Approximation.* Stephen J. Gardiner. London Math. Soc. Lect. Note Ser., V. 221. Cambridge Univ Pr, 1995, xiii + 132 pp, $32.95 (P). [ISBN 0-521-49799-X] Develops the theory to answer the following question: For which closed sets $E \subset \mathbb{R}^n$ can a function which is harmonic on a neighborhood of E be uniformly approximated on E by functions harmonic on all of $\mathbb{R}^n$? Includes applications to the Dirichlet problem and nonuniqueness of the Radon transform. BH

Geometry, T*(15: 1), L*. *Affine and Projective Geometry.* M.K. Bennett. Wiley, 1995, xvi + 229 pp, $44.95. [ISBN 0-471-11315-8] Well-written, concise coverage of the coordinatization of (Desarguesian) affine and projective space and the lattices of flats of these spaces; culminates with the Fundamental Theory of Projective Geometry. Appendices introduce field theory and Hilbert's noncommutative division ring. Chapters conclude with suggestions for further reading. JNC

Geometry, P, L**.** *Quasicrystals and Geometry.* Marjorie Senechal. Cambridge Univ Pr, 1995, xv + 286 pp, $59.95. [ISBN 0-521-37259-3] Results and conjectures of an intense first decade of research with a detailed account of the historical and scientific context. Focuses on relating the geometry of discrete point sets to the diffraction spectra of functions, and on the emerging theory of periodic tilings. Intermediate mathematical level. JNC

Operations Research, T(14: 1), C. *Elementary Linear Programming with Applications, Second Edition.* Bernard Kolman, Robert E. Beck. Comp. Sci. & Sci. Computing. Academic Pr, 1995, xxii + 449 pp, $59.95, with disk. [ISBN 0-12-417910-X] More linear algebra review, new exercises, an introduction to the Karmarkar algorithm, and a disk containing linear programming software SMPX. (1980 edition, TR, June–July 1980.) DH

Optimization, C, P. *Evolution and Optimum Seeking.* Hans-Paul Schwefel. Sixth-Generation Comp. Tech. Ser. Wiley. 1995, ix + 444 pp, $64.95, with disk. [ISBN 0-471-57148-2]

Optimization, T(16-17:1), P. *Stochastic Programming.* Peter Kall, Stein W. Wallace. Interscience Ser. in Systems & Optimization. Wiley, 1994, xii + 307 pp, $39.95 (P). [ISBN 0-471-95158-7]

Optimal Control, P. *Topics in the Calculus of Variations.* Martin Fuchs. Adv. Lect. in Math. Friedr Vieweg & Sohn, 1994, vii + 145 pp, $28 (P). [ISBN 3-528-06623-7]

Elementary Statistics, T(13: 1), C. *A First Course in Business Statistics, Sixth Edition.*

James T. McClave, P. George Benson, Terry Sincich. Prentice Hall, 1995, xx + 746 pp, with disk. [ISBN 0-02-379175-6] Many business applications, with emphasis on interpreting computer output. Answers given to all exercises. (1983 Dellen *Second Edition*, TR, November 1983.) DH

Elementary Statistics, T(13: 1), C. *A First Course in Statistics, Fifth Edition.* James T. McClave, Terry Sincich. Prentice Hall, 1995, xv + 608 pp, with disk. [ISBN 0-02-379195-0] New edition contains Poisson distribution, more exercises (including ones on interpreting p-values produced by software), and answers to all exercises. (1992 Dellen *Fourth Edition*, TR, May 1994.) DH

Computer Systems, P. *Essential System Administration, Second Edition.* Æleen Frisch. O'Reilly & Assoc, 1995, xvii + 758 pp, $32.95 (P). [ISBN 1-56592-127-5]

Computer Systems, P. *Applying RCS and SCCS.* Don Bolinger, Tan Bronson. O'Reilly & Assoc, 1995, xxiii + 501 pp, $29.95 (P). [ISBN 1-56592-117-8]

Computer Systems, P. *Using csh and tcsh.* Paul DuBois. O'Reilly & Assoc, 1995, xviii + 221 pp, $24.95 (P). [ISBN 1-56592-132-1]

Computer Systems, P. *Building Internet Firewalls.* D. Brent Chapman, Elizabeth D. Zwicky. O'Reilly & Assoc, 1995, xxvi + 517 pp, $29.95 (P). [ISBN 1-56592-124-0]

Computer Systems, C, P. *ORACLE PL/SQL Programming.* Steven Feuerstein. O'Reilly & Assoc, 1995, xxvii + 885 pp, $39.95 (P), with disk. [ISBN 1-56592-142-9]

Computer Science, S(15–18), P, L. *Logic and Representation.* Robert C. Moore. CSLI Lect. Notes, No. 39. Center for Study of Language & Information (Stanford U., Ventura Hall, Stanford, CA 94305), 1995, xiv + 196 pp, $19.95 (P); $39.95. [ISBN 1-881526-15-1; 1-881526-16-X] An edited collection of essays woven into a whole. The central point of view is on the role of formal logic and of explicit representation of knowledge in problems in artificial intelligence. Part 1 is on methodology; Part 2 deals with propositional attitudes; Part 3 presents autoepistemic logic; Part 4 addresses the semantics of natural language. RJA

Applications (Fluid Mechanics), S(18), P. *Solitons and Geometry.* S.P. Novikov. Cambridge Univ Pr, 1994, 58 pp, $19.95 (P). [ISBN 0-521-47196-6] Very sparse and highly technical. Unreadable by the non-expert. MU

Applications (Quantum Theory), S(18). *Linear Infinite-Particle Operators.* V.A. Malyshev, R.A. Minlos. Transl. of Math. Mono., V. 143. AMS, 1995, viii + 298 pp, $125. [ISBN 0-8218-0283-6] Studies the Moscow method and the direct cluster expansion method for spectral analysis, with most attention devoted to the latter. It builds upon the author's previous work, *Spectrum Analysis and Scattering Theory for a Three-particle Cluster Operator.* MU

Applications (Quantum Theory), T(18: 1, 2), S, P, L. *The Undivided Universe: An Ontological Interpretation of Quantum Theory.* D. Bohm, B.J. Hiley. Routledge, 1993, xii + 397 pp, $18.95 (P). [ISBN 0-415-12185-X] A thought-provoking philosophical excursion into physics, based upon an innovative approach to quantum theory that eliminates the assumed observer. The reader is expected to be very familiar with high-level physics. MU

Applications (Relativity), T(17–18: 1, 2), S, P, L. *The Geometry of Kerr Black Holes.* Barrett O'Neill. AK Peters, 1995, xvii + 381 pp, $79.95. [ISBN 1-56881-019-9] This text should delight mathematicians interested in general relativity. Begins with a brief historical overview and a careful tour of manifolds, tensors, and differential geometry. Most of the text is lucid high-level mathematics, well-motivated by the physics. MU

Applications (Relativity), T(16–17: 1, 2), S, P, L. *Gravitation and Inertia.* Ignazio Ciufolini, John Archibald Wheeler. Ser. in Physics. Princeton Univ Pr, 1995, xii + 498 pp, $49.50. [ISBN 0-691-03323-4] A rare achievement. Carefully chosen and expertly handled mathematics embedded in rich and lucid descriptive text. Readers are assumed to be familiar with the usual mathematics. MU

Applications (Systems Theory), P. *Lecture Notes in Control and Information Sciences–201: Robust Stability and Convexity: An Introduction.* Jacob Kogan. Springer-Verlag, 1995, xi + 176 pp, $41 (P). [ISBN 0-387-19919-5] Well worth a look—check out the following result (from page 3): "*Corollary*: Those who believe, or claim, that their method is the *best* one suffer from that alliterative affliction, ignorance/arrogance. *Proof*: By observation." BC

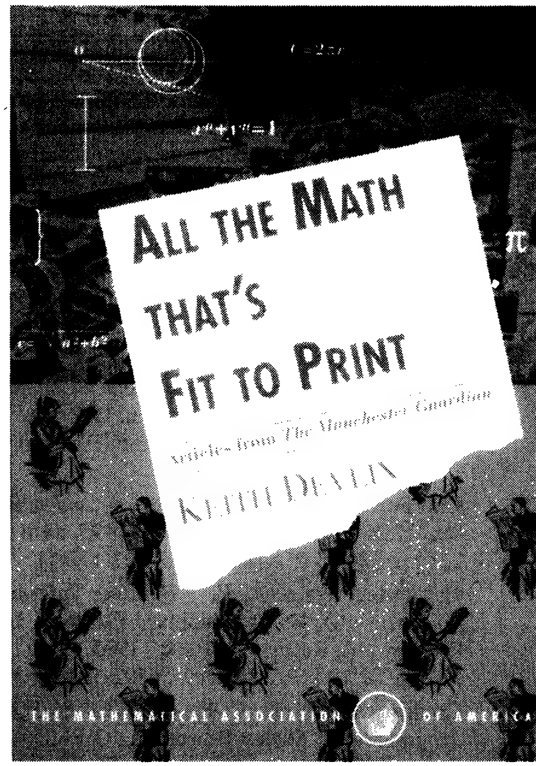
Reviewers

RJA: Richard J. Allen, St. Olaf; JNC: Judith N. Cederberg, St. Olaf; BC: Barry Cipra, St. Olaf; CEC: Clifton E. Corzatt, St. Olaf; PF: Paul Froeschl, Macalester; BH: Bruce Hanson, St. Olaf; DH: Deanna Haunsperger, Carleton; SK: Steve Kennedy, Carleton; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; JCS: Janice Sklensky, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; MU: Milton Ulmer, Carleton.

All the Math That's Fit to Print

Articles from the Manchester Guardian

Keith Devlin



Mathematics and mathematicians can be the objects of public interest, if there are individuals capable of explaining those items in a form that the intelligent reader can follow. Keith Devlin is such a person and the editors of the British paper, The Manchester Guardian, were intelligent enough to understand that. This book should be an element of every public library.

—Journal of Recreational Mathematics

This new work reveals another side of Devlin's interesting investigations into mathematics and his efforts to share them with laypersons...Anyone interested in mathematics will find something of interest in this book...When possible, the author provides a historical context for the new ideas being explored.

—Choice

Between 1983 and 1989 Keith Devlin, research mathematician, author and educator, wrote a semi-monthly column on mathematics and computing in the English national daily newspaper, The Manchester Guardian. This book is a compilation of many of those articles. It is a witty, entertaining, easy-to-read piece of work.

The mathematical topics range from simple puzzles to deep results including open problems such as Faltings Theorem and the Riemann Conjecture. You will find articles on prime numbers, how to work out claims for traveling expenses, calculating pi, computer simulation, patterns and palindromes, cryptology, and much more.

This book is meant for browsing by anyone who regularly reads a serious newspaper and has some interest in matters scientific or mathematical.

345 pp., Paperbound, 1994
ISBN - 0-88385-515-1
List: $38.50
MAA Member: $28.95
Catalog Code: ATMA/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	ATMA/JR	______	______
______	______________	______	______
		TOTAL	______

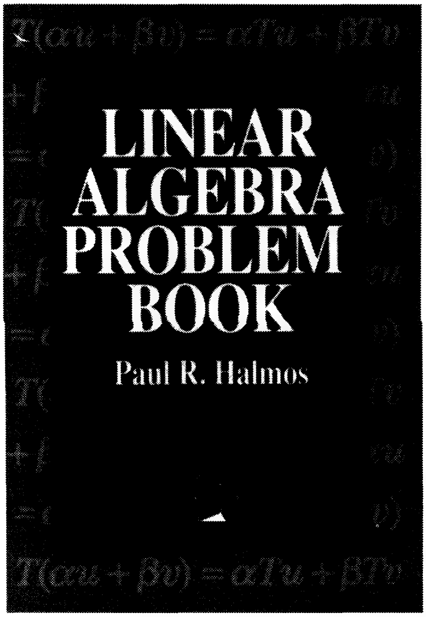
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________________ Expires ___/___

Signature ______________________

Linear Algebra Problem Book

Paul R. Halmos



Were it possible for the experience of apprenticeship to a master of mathematics to be packaged between the covers of a book, this would be it. No teacher of linear algebra should neglect to consult it. Highly recommended for all libraries.
– Choice Magazine

This is a book for mathematicians at all levels. Paul Halmos tells us, "Even if I know some answers, I don't think I understand a subject until I know the questions. The questions in mathematics are called problems—and although I learned some linear algebra a long time ago, until now I have made no serious effort to examine the problems that the solutions are based on. I wrote this book to organize those questions—problems—in my own mind."

This book is useful to anyone who needs linear algebra—and nowadays that means every user of mathematics. It can be used as the basis of either an official course or a program of private study.

If used as a course, the book can stand by itself, or if so desired, it can be stirred in with a standard linear algebra course as the seasoning that provides the interest, the challenge, the motivation that is needed by experienced scholars as much as by beginning students.

Contents

340 pp., Paperbound, 1995
ISBN 0-88385-322-1
List: $39.95
MAA Member: $28.95
Catalog Code: DOL-16/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________________

Address ______________________________

City ______________________________

State ______________ Zip ______________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-16/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

Five Hundred Mathematical Challenges

Edward J. Barbeau, William O. Moser, and Murray S. Klamkin

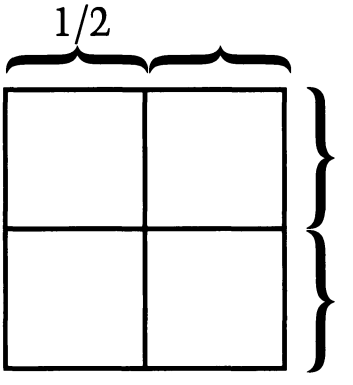
This book contains 500 problems that range over a wide spectrum of areas of high school mathematics and levels of difficulty. Some are simple mathematical puzzlers while others are serious problems at the Olympiad level. Students of all levels of interest and ability will be entertained and taught by the book. For many problems, more than one solution is supplied so that students can see how different approaches can be taken to a problem and compare the elegance and efficiency of different tools that might be applied.

Teachers at both the college and secondary levels will find the book useful, both for encouraging their students and for their own pleasure. Some of the problems can be used to provide a little spice in the regular curriculum by demonstrating the power of very basic techniques.

These problems were first published as a series of problem booklets almost twenty years ago, at a time when there were few resources of this type available for the English reader. They have stood the test of time and the demand for them has been steady. Their publication in book form is long overdue.

This collection provides a solid base for students who wish to enter competitions at the Olympiad level. They can begin with easy problems and progress to more demanding ones. A special mathematical tool chest summarizes the results and techniques needed by competition-level students.

Problem 14.
Show that if 5 points are all in, or on, a square of side 1, then some pair of them will be no further than $\frac{\sqrt{2}}{2}$ apart.



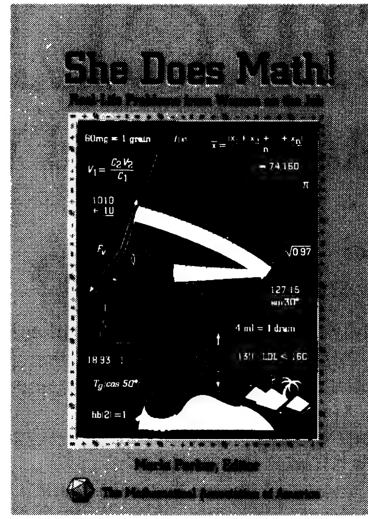
Solution.
Divide the square into four squares of side 1/2. By the Pigeonhole Principle, one of these four squares contains at least two of the points, whose distance apart must be no greater than the diagonal of the square of side 1/2, namely $\frac{\sqrt{2}}{2}$.

300 pp., Paperbound, 1995
ISBN 0-88385-519-4
List: $29.50
MAA Member: $24.95
Catalog Code: CHMP/JR

She Does Math!

Real-Life Problems From Women On The Job

Marla Parker, Editor



She Does Math! presents the career histories of 38 professional women and math problems related to their work. Each history describes how much math the author took in high school and college; how she chose her field of study; and how she ended up in her current job. Each of the women presents problems that are typical of those she has faced in her job. The problems require only high school mathematics for their solution.

Who should have this book?

Your daughter, your granddaughter, your sister, your former math teacher, your students—and young men too. They want to know how the math they study is applied, and this book will show them.

There are many reasons to buy this book:

➤ By reading the career histories of the women profiled in this book, young people will learn that if they take mathematics courses in high school and college they will be qualified to enter interesting technical fields and earn good salaries.
➤ The problems have special appeal to students who are beginning to think about career choices.
➤ The book provides practical information about the job market in an interesting, innovative way.
➤ Strong female role models who work as successful technical professionals are presented.
➤ The problems are interesting and challenging, yet require only high school mathematics. They demonstrate how good math skills are applied to real-life problems.

Read what others have said about She Does Math!

Finally — a practical, innovative, well-written book that will also inspire its readers. The wonder is...it's a mathematics text and a biography! The idea of women telling their own career stories, emphasizing the mathematics they use in their jobs is extremely creative. This book makes me wish that I could go through school all over again!
Anne L. Bryant, Executive Director
American Association of University Women

She Does Math! *will undoubtedly appeal both to students who already enjoy math and want to get a view of potential career paths, and also to students who want to better understand the relevance of their math classes to their future careers. It is an absorbing look into the lives of some very inspiring and talented women!*
Susanne Hupfer and Elisabeth Freeman
Yale University

This collection is a wonderful confirmation that real women do math. They do math in a surprising variety of careers, fully enjoying the challenge and rewards of solving complex problems. This is a book for young women and men, a book for their teachers and parents, a book that informs about the possibilities that mathematics affords to all. It is also a book that will engage you in real-life mathematics! – Doris Schattschneider
Moravian College

272 pp., Paperbound, 1995 ISBN 0-88385-702-2
List: $27.50 MAA Member: $20.95
Catalog Code: SDM/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

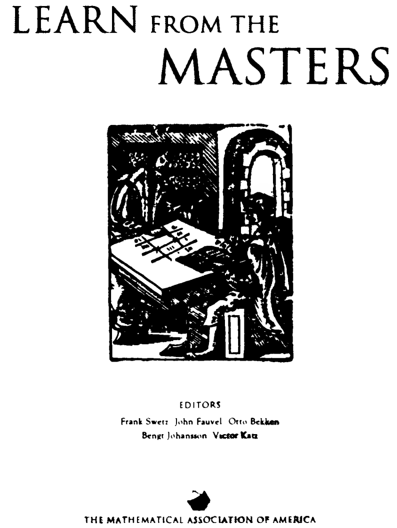
QTY.	CATALOG CODE	PRICE	AMOUNT
_____	SDM/JR	_____	_____
_____	_____	_____	_____
		TOTAL	_____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

Learn from the Masters



Frank Swetz, John Fauvel, Otto Bekken, Bengt Johansson, Victor Katz, Editors

Provides high school and college teachers with important historical ideas and insights which can be immediately applied in the classroom.

This book is for college and high school teachers who want to know how they can use the history of mathematics as a pedagogical tool to help their students construct their own knowledge of mathematics. Often, a historical development of a particular topic is the best way to present a mathematical topic, but teachers may not have the time to do the research needed to present the material. This book provides its readers with historical ideas and insights which can be immediately applied in the classroom.

The book is divided into two sections: the first on the use of history in high school mathematics, and the second on its use in university mathematics. So, high school teachers planning a discussion of logarithms, will find here the historical background of that idea along with suggestions for incorporating that history in the development of the idea in class. College teachers of abstract algebra will benefit by reading the three articles in the book dealing with aspects of that subject and considering their ideas for presenting groups, rings, and fields.

The articles are diverse, covering fields such as trigonometry, mathematical modeling, calculus, linear algebra, vector analysis, and celestial mechanics. Also included are articles of a somewhat philosophical nature, which give general ideas on why history should be used in teaching and how it can be used in various special kinds of courses. Each article contains a bibliography to guide the reader to further reading on the subject.

This book grew out of a conference in Norway which brought together mathematicians and mathematics educators from a dozen countries who were interested in the use of the history of mathematics as a pedagogical tool in the teaching of mathematics. Since the conference which provided the genesis of this book took place in Norway near the home where Niels Henrik Abel spent his final days, the book's title comes from a note scribbled in one of Abel's notebooks: "It appears to me that if one wants to make progress in mathematics one should study the masters." The authors hope that readers will benefit from Abel's advice and show their students how they too can Learn from the Masters.

312 pp., Paperbound, 1995
ISBN-0-88385-703-0
List: $26.50
MAA Member: $20.95
Catalog Code: LRM/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	LRM/JR	______	______
______	______	______	______
		TOTAL	______

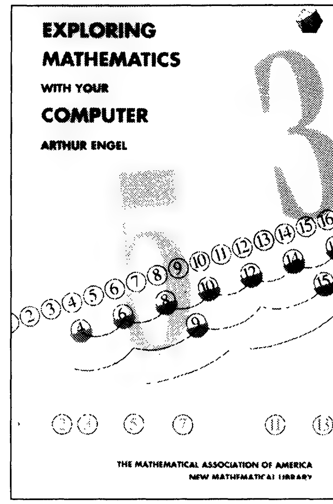
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

Exploring Mathematics With Your Computer

Arthur Engel



Selected by Choice magazine as one of the outstanding academic books of 1994.

A must for academic libraries supporting an undergraduate major in mathematics. Public libraries with strong science collections should have this book as a resource for traditional mathematics topics and as recreation for the mathematically inclined. Capable high school students would also benefit.

— Choice

This book...is intended for students, mathematics teachers and mathematicians who are just starting to explore mathematics on their own computers. In studying it, and especially in working through its exercises, readers will get to know many new, elementary topics, and they will learn as much from the extensive exercises as from the examples.

— Mathematical Reviews

Those who have met examples of Arthur Engel's work will know what to expect, and this is another example of his astonishing versatility...When Engel's Elementarmathematik von algorithmischen Standpunkt first appeared, Trevor Fletcher is reputed to have said something like "Take care—it will change your life!: Much the same is true of this book. Buy it!

—The Mathematical Gazette

The author's style is pleasant, clear and precise and his choice of material is superb. The book is a gold mine, and I will return to it many times to dig out more nuggets. Congratulations, Professor, Engel, for one of the best texts I have seen. It is highly recommended.

— SPECTRUM

This is a mathematics book, not a programming book, although it explains Pascal to beginners. It is aimed at high school students and undergraduates with a strong interest in mathematics and teachers looking for fresh ideas. It is full of diverse mathematical ideas requiring little background. It includes a large number of challenging problems, many of which illustrate how numerical computation leads to conjectures which can then be proved by mathematical reasoning.

It is assumed that readers have a PC at their disposal.

You need a Pascal compiler in order to use the 3.5″ IBM-compatible disk packaged with this volume.

264 pp., Paperbound, 1993
ISBN 0-88385-636-0
List: $45.50
MAA Member: $32.50
Catalog Code: NML-35/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (202) 387-5200 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________________

Address ______________________________

City ______________________________

State ______________ Zip ______________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	NML-35/JR	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 103, Number 5 May 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428B
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

What Is a Surface?

Frank Morgan

A search for a good definition of surface leads to the rectifiable currents of geometric measure theory, with interesting advantages and disadvantages.

For details and references see Morgan's *Geometric Measure Theory: a Beginner's Guide*, Academic Press, 2nd edition, 1995. Figures by James F. Bredt.

.. **WHAT IS AN INCLUSIVE DEFINITION OF A GENERAL SURFACE IN R^3?** We want to include smooth embedded manifolds with boundary, as in Figure 1, and we want to be able to allow singularities, as in the cube and cone of Figure 2.

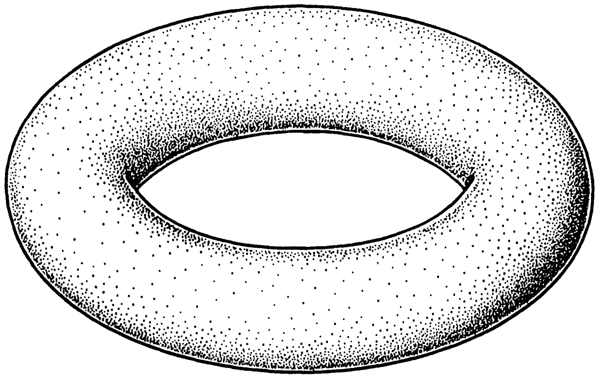


Figure 1. The nicest surfaces are smooth, embedded manifolds.

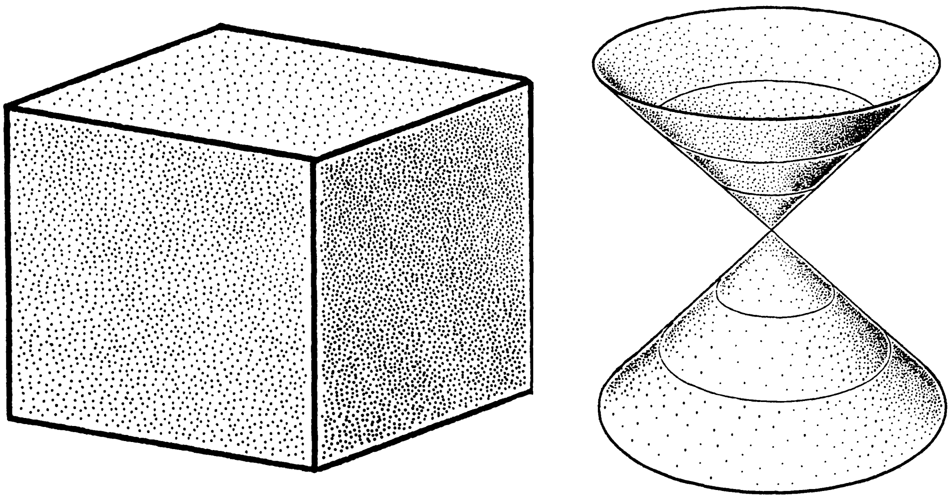


Figure 2. Surfaces such as the cube and cone have mild singularities.

We might allow any smooth embedded *stratified manifold*, i.e., a set that is a smooth embedded 2-dimensional manifold, except for a subset that consists of smooth embedded curves, except for a set of isolated points. We might go farther and allow any set that is a smooth embedded 2-dimensional manifold except for a set of 2-dimensional measure 0.

Unfortunately for such surfaces it is hard to prove the existence of area-minimizers or solutions to other geometric variational problems, because such classes of surfaces are not closed under the limit arguments used to obtain such solutions. What good is a sequence of surfaces with areas approaching an infimum, if there is no limit surface realizing the least area? Moreover, there are more general sets that deserve to be called surfaces. We begin with a lower-dimensional example of a compact, connected 1-dimensional "curve" in $\mathbf{R}^2$ that is not a stratified manifold, and actually has a singular set of positive 1-dimensional measure. The construction is based on a Cantor set C in $\mathbf{R}^1$ of positive measure, differing from the usual Cantor set in the rapidly diminishing size of the segments removed.

2. CANTOR SET C OF POSITIVE MEASURE. To construct a Cantor set C of positive measure, start with the unit interval as in Figure 3a. Remove the middle $1/4$. From each of the remaining two pieces, remove an open segment in the middle of length $1/16$. At the kth step, remove from each of the remaining 2^k pieces an open segment in the middle of length $\frac{1}{4} \cdot 2^{-2k}$. In countably infinitely many steps you remove total length

$$\frac{1}{4} \sum_{k=0}^{\infty} 2^k 2^{-2k} = \frac{1}{2},$$

leaving a compact Cantor set C of length $1/2$ as in Figure 3b.

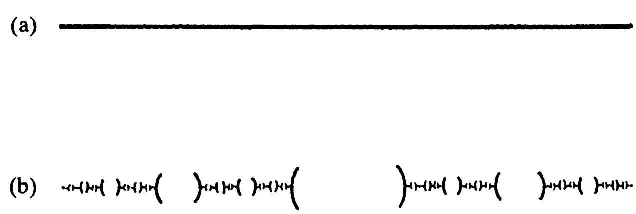


Figure 3. Start with the unit interval. Removing infinitely many open segments of rapidly decreasing length leaves a Cantor set C of positive measure.

3. A "CURVE" WITH A SINGULAR SET OF POSITIVE MEASURE. We now construct a curve that is singular on our Cantor set C of positive measure. Let A consist of a nice smooth path from $(0,0)$ to $(1,0)$ above the x-axis, together with its reflection below the x-axis, as in Figure 4a. Replace every segment removed in the previous construction of the Cantor set C by a suitably scaled copy of A, as in Figure 4b. The resulting "curve" is an embedded manifold except for the singular set C, which has 1-dimensional measure $1/2$.

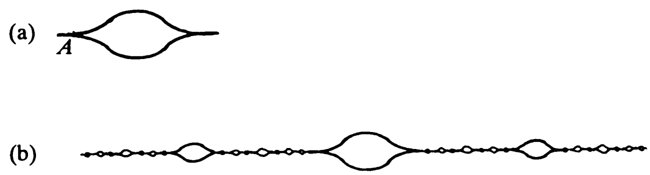


Figure 4. Inserting bifurcating paths A into the gaps of a Cantor set C of positive measure yields a "curve" with a singular set of positive 1-dimensional measure.

Similarly, one could add infinitely many handles of finite total area to the sphere to produce a 2-dimensional surface S with a singular Cantor set of positive 2-dimensional measure, as in Figure 5. Note that S is the limit of smooth submanifolds without boundary, namely, spheres with a large finite number of handles.

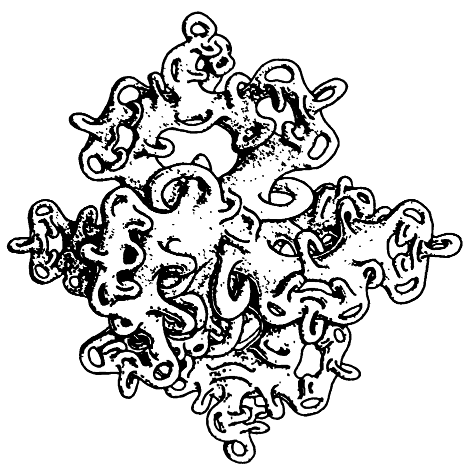


Figure 5. Adding infinitely many handles of finite total area to the sphere can produce a surface S with a singular Cantor set of positive 2-dimensional measure.

4. RECTIFIABLE SETS. A good class of 2-dimensional subsets of $\mathbf{R}^3$ that includes all of the surfaces we have considered or would ever want to consider and has nice closure properties under limit operations is the *rectifiable sets* of H. Federer. What do they include? To begin with, they include the image of any reasonable function f from planar domains into $\mathbf{R}^3$. The defining function f is not required to be differentiable, merely *Lipschitz*:

$$|f(x) - f(y)| \leq C|x - y|, \tag{1}$$

allowing for example the upper cone as the image of

$$\begin{aligned} f\colon \mathbf{R}^2 &\to \mathbf{R}^3 \\ x &\mapsto (x, |x|). \end{aligned}$$

This Lipschitz condition (1), by bounding the amount of stretching, is just right for producing surfaces of finite area. To obtain the full class of rectifiable sets, allow arbitrary (measurable) subsets of countable unions of such images of Lipschitz

functions, as long as the total area remains finite. Such rectifiable sets include our example of a sphere with infinitely many handles.

It is a very fortunate theorem of real analysis that Lipschitz functions are differentiable almost everywhere. As a result, although rectifiable sets can be quite intricate, they turn out to have a kind of measure-theoretic "approximate" tangent plane at almost every point, which is good enough for doing lots of geometry. For example, one can define an orientation of a rectifiable set simply as a measurable orientation of (almost every) tangent plane. So far there is no coherence from point to point, and a smooth piece of surface has infinitely many "orientations." We will see that incoherent orientations can be detected by the extra boundary they introduce.

5. THE STRUCTURE THEORY. The fundamental role of rectifiable sets is exposed by a general structure theorem of Besicovitch and Federer. The theorem says that every subset E of $\mathbf{R}^3$ of finite 2-dimensional measure can be decomposed as a rectifiable set and a *purely unrectifiable set* that is invisible from almost all directions (whose projections onto almost all planes have measure 0). So rectifiable sets form a fundamental and inclusive class of surfaces. The question is whether we can do geometry with them.

6. HAUSDORFF METRIC UNSUITABLE. How can you say when two rectifiable sets are close together? The standard Hausdorff metric distance between two compact subsets of $\mathbf{R}^3$ is defined as the greatest distance between any point of one and any point of the other. Rectifiable sets need not be compact, and the Hausdorff metric is practically useless, as shown by the following example of radically different rectifiable sets S_0 and S_1 that are close together in the Hausdorff metric. Let S_0 be the unit sphere. Let S_1 be a countable collection of tiny spheres of radius at most ε, centered in S_0, dense in S_0, with total area ε. Then the Hausdorff distance between S_0 and S_1 is at most ε, even though S_0 is the round sphere of area 4π, and S_1 is a fragmented (though in some sense boundaryless) set of area at most ε.

7. RECTIFIABLE CURRENTS. So how can we apply geometric concepts to rectifiable sets? The way to define boundary and topology for oriented rectifiable sets is to view them as *currents*, linear functionals on smooth differential forms φ. Since an oriented rectifiable set S has an approximate oriented tangent plane $\vec{S}$ at almost every point, one can integrate a differential form φ over S:

$$S(\varphi) = \int_S \varphi(\vec{S}),$$

and thus view S as a current. The space of currents so arising from rectifiable sets is called the space $\mathscr{R}_2\mathbf{R}^3$ of 2-dimensional rectifiable currents. One allows integral multiplicities, but finite total area and compact support.

The general concept of currents was a generalization, due to G. de Rham, of distributions. H. Federer and W. Fleming introduced rectifiable currents in 1960 in their foundational paper on geometric measure theory, which won the AMS Steele Prize for its fundamental importance.

8. BOUNDARY. The boundary of a current S, denoted ∂S, may be defined as an abstract current by the formula

$$\partial S(\varphi) = S(d\varphi).$$

By Stokes's theorem, this definition agrees with the usual one for smooth oriented manifolds with boundary. Of course there is no reason in general that the boundary ∂S of a rectifiable current S should be a rectifiable current. If it happens to be, then the original current S is called an *integral current*.

The boundary of the unit disc with the standard orientation is the unit circle with the standard orientation, and therefore the unit disc is a nice integral current. Suppose however the disc is decomposed into the infinitely many concentric annuli

$$A_n = \{1/(n+1) < r \leq 1/n\}$$

of Figure 6 and they are given alternating orientations. Then the boundary includes infinitely many circles (with multiplicity 2), has infinite length, hence is not a rectifiable current, and therefore the disc with this incoherent orientation is not an integral current. This example shows how incoherent orientations may be detected by the additional boundary they introduce.

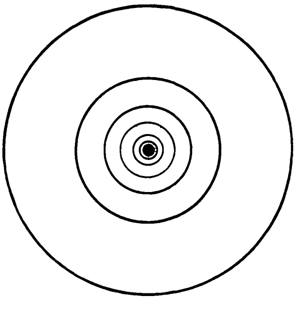


Figure 6. Giving alternating orientations to concentric annuli creates lots of extra boundary.

9. TOPOLOGY AND FLAT NORM. The notion of boundary leads to a topology on the space of rectifiable currents given by H. Whitney's *flat norm* $\mathscr{F}$ on currents, defined by

$$\mathscr{F}(S) = \inf\{\text{area } T + \text{vol } R \colon S - T = \partial R\}.$$

For example, the two discs D_1, D_2 of Figure 7 are close together in this topology because their difference $S = D_2 - D_1$ together with a thin band T bounds a region R of small volume. Whitney wanted to distinguish the flat norm from another larger sharp norm: as a music major, he borrowed the terms "flat" and "sharp," indicating lower or higher notes, from musical terminology.

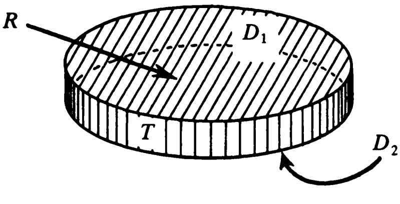


Figure 7. The two discs D_1, D_2 are close together in flat norm $\mathscr{F}$ because their difference $S = D_2 - D_1$ together with a thin band T bounds a region R of small volume.

10. THE COMPACTNESS THEOREM. The almost miraculous payoff from the notions of boundary and topology from currents is Fleming's compactness theorem inside a ball B in $\mathbf{R}^3$:

$$\{\text{Rectifiable currents } S \text{ in } B\colon \text{area } S \leq c, \text{ length } \partial S \leq c\}$$

is compact under the flat norm $\mathcal{F}$. In other words, any infinite sequence of our rectifiable surfaces in the room you are sitting in, with bounds on the area and boundary length, has a convergent subsequence. Consequently, reasonable geometric variational problems have solutions, as we will now illustrate with area-minimizing surfaces. No such compactness holds for smooth submanifolds.

11. EXISTENCE OF AREA-MINIMIZING SURFACES. *Let C be a closed bounded rectifiable curve of any number of components in* $\mathbf{R}^3$. *Then C bounds a rectifiable current of least area.*

Proof: The curve C lies in some large ball B about 0. It is easy to show C bounds some rectifiable current, for example the cone over C as in Figure 8. Let S_i be a sequence of rectifiable currents bounded by C with areas converging to the infimum. By projecting the surfaces back into B if necessary, which does not increase area, we may assume all the S_i lie in B. By the compactness theorem, we may assume the S_i converge to some limit S. It is easy to show that area $S \leq \mathit{lim}$ area S_i, and hence S provides an area-minimizing surface.

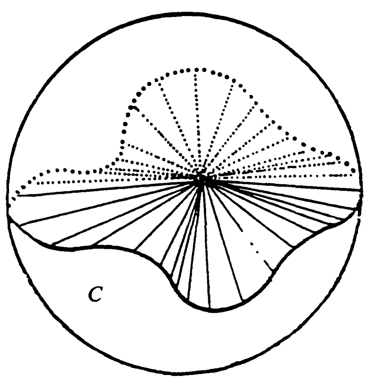


Figure 8. Any closed rectifiable curve C bounds some surface, for example the cone over C.

12. THE REGULARITY THEOREM. For a given boundary curve, we now have an area minimizer S in the class of rectifiable currents. The big remaining question is whether S is a reasonable surface. The answer due to Fleming (1962) and R. Hardt and L. Simon (1979) sounds too good to be true:

An area-minimizing surface (rectifiable current) bounded by a smooth curve in $\mathbf{R}^3$ *is a smooth submanifold with boundary.*

Thus it turns out that although we allowed all kinds of singularities, area-minimizing rectifiable currents do not have any. Such complete regularity fails for other classes of surfaces, such as classical mappings of the disc. For the boundary pictured in Figure 9, a circle with a tail, the area-minimizing disc passes through

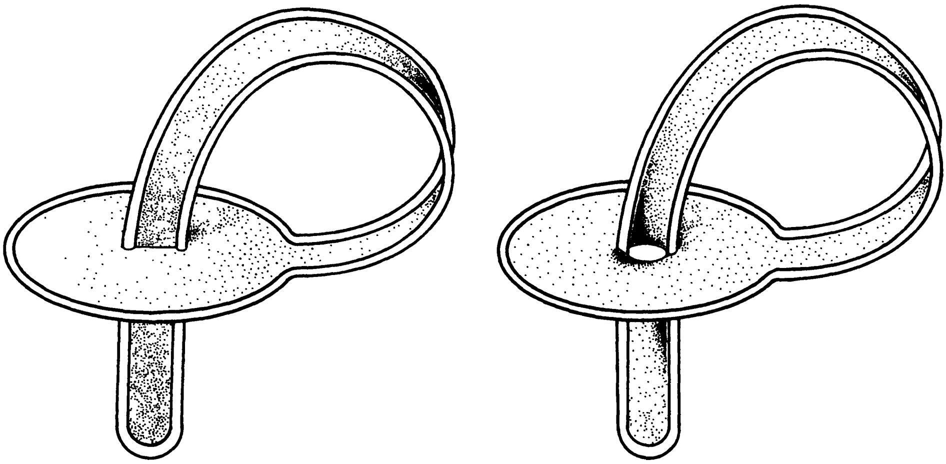


Figure 9. A classical area-minimizing disc need not be embedded, but the area-minimizing rectifiable current S always is embedded. Here S has higher genus and less area than the disc.

itself. The area-minimizing rectifiable current has higher genus, has less area, and is embedded. It flows from the top, flows down the tail, pans out in back onto the disc, flows around front, and flows down the tail to the bottom. There is a hole in the middle that you can stick your finger through. Incidentally, this surface exists as a soap film, whereas the least-area disc does not. If the boundary curve is badly knotted, it bounds no embedded disc, and the area-minimizing surface will necessarily have high genus. The presumed area-minimizing (orientable) surface bounded by a trefoil knot is pictured in Figure 10.

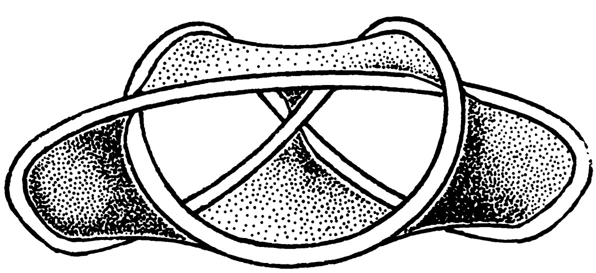


Figure 10. The area-minimizing (orientable) surface bounded by this trefoil knot is presumably this embedded surface of genus 1.

13. HIGHER DIMENSIONS. The theory of rectifiable currents generalizes to m-dimensional surfaces in $\mathbf{R}^n$. Area-minimizing hypersurfaces remain smooth submanifolds through $\mathbf{R}^7$; for $n \geq 8$, area-minimizing hypersurfaces in $\mathbf{R}^n$ can have $(n - 8)$-dimensional singular sets. In higher codimension, singularities occur, even for 2-dimensional surfaces in $\mathbf{R}^4$.

14. FAILINGS OF RECTIFIABLE CURRENTS. For all their virtues, rectifiable currents have their failings too. Rectifiable currents must be oriented, while many physical surfaces need not be oriented. W. Ziemer's *rectifiable currents modulo two* yield a similar theory of unoriented surfaces.

Figure 11. Bill Ziemer (right), who introduced flat chains modulo 2, with his thesis advisor, Wendell Fleming (left), and the author (center), at a celebration in Ziemer's honor at Indiana in 1994. Photo courtesy of Ziemer.

Physical surfaces such as soap films often consist of pieces of surface meeting along whole singular curves. These curves, although not part of the given boundary, unfortunately count as boundary for rectifiable currents. Explaining the structure of soap films required a new theory of $(\mathbf{M}, \varepsilon, \delta)$-*minimal sets* developed by F. Almgren and J. Taylor.

Crystal surfaces often exhibit an infinitesimal corrugation well modeled by the *varifolds* of Almgren and W. Allard.

15. OPEN QUESTIONS. There are many fundamental open questions. For example, for the existence of area-minimizing surfaces in $\mathbf{R}^3$, is there a simple direct proof that stays inside the class of smooth submanifolds? Are area-minimizing surfaces in general dimensions stratified manifolds, or can they have fractal singular sets?

Department of Mathematics
Williams College
Williamstown, MA 01267
Frank.Morgan@williams.edu

Non-Euclidean Flashlights

Ron Perline

The fact that the parabola reflects parallel lines to a single point was known to antiquity. We make good use of the fact every day: parabolic mirrors are found everywhere from flashlights to reflector telescopes. As it happens, the local geometry of our world is (to a good approximation) euclidean; if this were not the case, a different reflecting shape would be necessary. So how do you build a non-euclidean flashlight?

We will answer the question in the case that the geometry in question is hyperbolic, with the Poincaré half-plane as model. The Poincaré half-plane is a famous example from differential geometry [B], but we need only three facts concerning its geometric structure:

(i) Its geodesics (which play the role of light rays or straight lines in euclidean space) are vertical half-lines, or half-circles with centers on the x-axis (see Figure 1).

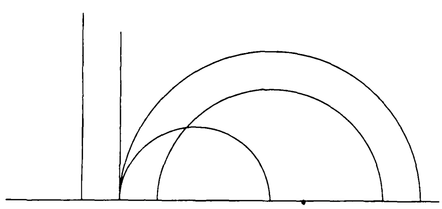


Figure 1

(ii) Hyperbolic geometry is *conformal* to euclidean geometry. For our purposes, this simply means that any argument concerning a configuration of points and curves in euclidean space, which involves only angles and ignores distances, is valid without change in hyperbolic geometry.

(iii) The law of reflection—when light rays encounter a reflecting curve, the angle of incidence equals the angle of reflection—holds true in hyperbolic space.

We will assume that the reflecting boundary can be represented as a graph $y = f(x)$, and ask that the reflecting curve focus the pencil of vertical parallel rays onto a single point. We will translate this geometric property into a differential equation for f. It turns out that the relevant differential equation is an elegant and classical one; we review a method of solution. With solution in hand, we take a look at a few of these reflecting curves. Finally, we discuss connections with recent developments in the theory of integrable systems.

TWO AUXILIARY FORMULA. In order to derive our differential equation, we need two formulae from elementary analytic geometry. First, consider the setup of Figure 2. We have the graph of some function $f(x)$ in the plane, with tangent line

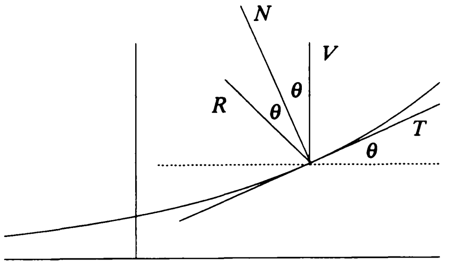


Figure 2

T. T forms an angle θ with the x-axis, and of course we have $f'(x)$ = slope of $T = \tan\theta$. If V is a vertical ray, N the normal ray, and R the ray that results from reflecting V off the curve $y = f(x)$, then it is easy to see that

$$m_R = \text{slope of } R = \tan(\pi/2 + 2\theta) = -\cot(2\theta)$$
$$= \frac{\tan^2\theta - 1}{2\tan\theta} = \frac{(f'(x))^2 - 1}{2f'(x)}.$$

This formula was derived by considering the angular relations in Figure 2. We now use the fact the hyperbolic geometry is conformal to euclidean geometry, which allows us to argue that, in the hyperbolic case, the slope of the tangent line of the reflected circular arc is also $((f'(x))^2 - 1)/(2f'(x))$ (see Figure 3).

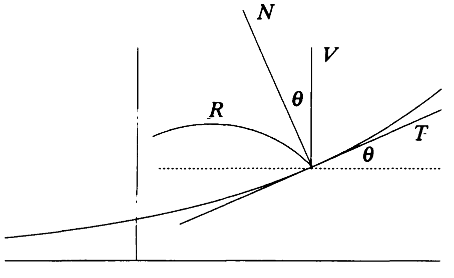


Figure 3

The second formula that we will need is the hyperbolic analogue of the point-slope formula of ordinary euclidean geometry. Let $p = (x_0, y_0)$ be a point in the upper half-plane, and let m be some real value. We want the equation of the semicircle with center on the x-axis, which passes through p, and whose tangent line at p has slope m (see Figure 4).

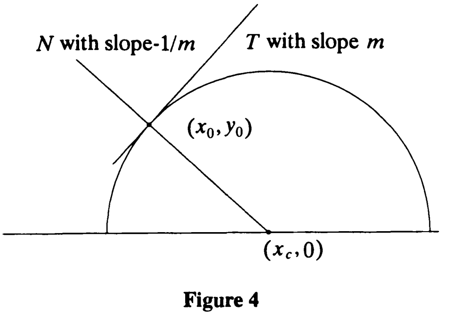


Figure 4

The normal line clearly has equation $y - y_0 = -(1/m)(x - x_0)$, which intersects the x-axis at the point $x_c = my_0 + x_0$, which is the location of the center of the desired semicircle. The radius satisfies $R^2 = (x_0 - x_c)^2 + y_0^2 = (m^2 + 1)y_0^2$. Thus the equation of the desired semicircle is $(x - x_c)^2 + y^2 = (m^2 + 1)y_0^2$. For future reference, we observe that this last line easily implies that y_I, the y-intercept of the semicircle (if it has one) satisfies $y_I^2 = y_0^2 - 2mx_0y_0 - x_0^2$.

DERIVATION OF THE DIFFERENTIAL EQUATION. Let us now consider a curve in the upper half-plane, described by the graph of the function $y = f(x)$. If $(x_0, y_0) = (x_0, f(x_0))$ is a point on the curve, then the first of our auxiliary formulae tells us that an incoming vertical ray reflects off the curve along a semicircular arc with tangent line whose slope is $m_R = ((f'(x_0))^2 - 1)/(2f'(x_0))$. Our second formula allows us to conclude that y_I satisfies

$$y_I^2 = y_0^2 - 2mx_0y_0 - x_0^2 = f(x_0)^2 + \left((f'(x_0))^2 - 1\right)/(f'(x_0))(x_0f(x_0)) - x_0^2.$$

The discussion in the preceding paragraph holds for any curve. Suppose that f has the property that it reflects the pencil of vertical parallel lines to a single point; and suppose for definiteness that this focal point lies on the y-axis. This condition is equivalent to y_I being a constant independent of x_0, or that f satisfies

$$c^2 = f(x)^2 + (1/f'(x))\left(1 - f'(x)^2\right)(xf(x)) - x^2.$$

This equation can be rewritten as

$$c^2/xf + x/f - f/x = 1/f' - f'.$$

We make a slight change of notation (replace $f(x)$ by $y(x)$), and face the task of solving the differential equation

$$c^2/xy + x/y - y/x = 1/y' - y'. \tag{1}$$

THE CONFOCAL EQUATION. An interesting application of elementary differential equations to geometry is the construction of families of *orthogonal curves* or *orthogonal trajectories*. ([**S-R**], p. 117). We recall the basic notion: if we consider the solutions to a differential equation $y' = f(x, y)$, whose graphs in general form a one-parameter family of curves that fill out the plane, then the solution curves for the related differential equation $y' = -1/f(x, y)$ meet the solution curves to the first differential equation at ninety degree angles. Another way of stating the relation between the two equations is that the second equation is obtained from the first by replacing y' by $-1/y'$.

When we perform such a replacement in (1), we find that we obtain the *same* equation: (1) is *invariant* with respect to our y' transformation. At first glance we are led to the seemingly paradoxical notion that the solution curves to (1) are somehow "self-orthogonal". A closer look at (1) indicates the resolution of this "paradox": the equation is not of the form $y' = f(x, y)$, but rather (after multiplication by y') it is a quadratic expression in y'. One can then solve for y' in the resulting expression, obtaining two equations $y' = f_1(x, y)$ and $y' = f_2(x, y)$, whose solution curves form orthogonal families. We do not bother writing down the expressions for f_1 and f_2 because they are not particularly useful to us. Nevertheless, it now seems plausible that, to find solutions of (1), we should work with ideas related to the geometry of orthogonal curves. This turns out to be correct, and leads us to the study of the geometry of conics that are *confocal* (share the same foci).

Consider the equation ([**S-R**, p. 119, problem 5])

$$x^2/(a^2+\lambda)+y^2/(b^2+\lambda)=1 \text{ where } b^2>a^2>0 \text{ are fixed.} \tag{2}$$

This equation implicitly defines λ as a function of x and y; in fact it defines two such functions, one with values ranging from $-b^2$ to $-a^2$, the other with values ranging from $-a^2$ to ∞. We note that the level curves of λ form families of ellipses and hyperbolas, which are in fact confocal. The equation $\lambda(x, y) = k$ implicitly defines y as a function of x. We now derive a differential equation that has y as its solution.

Implicitly differentiating (2) with respect to x and y respectively, we obtain

$$\frac{2x}{a^2+\lambda}=\lambda_x\left(\frac{x^2}{(a^2+\lambda)^2}+\frac{y^2}{(b^2+\lambda)^2}\right)$$

and

$$\frac{2y}{a^2+\lambda}=\lambda_y\left(\frac{x^2}{(a^2+\lambda)^2}+\frac{y^2}{(b^2+\lambda)^2}\right).$$

Thus $\vec{n}=\operatorname{grad}\lambda=(\lambda_x,\lambda_y)$ is pointwise proportional to $(x/(a^2+\lambda), y/(b^2+\lambda))$. It follows that the vector $(-y/(b^2+\lambda), x/(a^2+\lambda))$ is tangent to the curve $\lambda = k$. The ratio $-(x/y)(b^2+\lambda)/(a^2+\lambda)$ is thus equal to $y'(x)$, where again y is defined implicitly by $\lambda = k$.

As is standard in such calculations, we want to get rid of the λ dependence. From the previous paragraph, we obtain

$$-y'=(x/y)(b^2+\lambda)/(a^2+\lambda) \text{ and } 1/y'=-(y/x)(a^2+\lambda)/(b^2+\lambda).$$

If we add and simplify, the result is

$$\begin{aligned}
1/y'-y' &= -(y/x)(a^2+\lambda)/(b^2+\lambda)+(x/y)(b^2+\lambda)/(a^2+\lambda)\\
&= -(y/x)(b^2+\lambda+a^2-b^2)/(b^2+\lambda)\\
&\quad +(x/y)(a^2+\lambda+b^2-a^2)/(a^2+\lambda)\\
&= -y/x+x/y+(b^2-a^2)\\
&\quad \times\left[(x/y)(1/(a^2+\lambda))+(y/x)(1/(b^2+\lambda))\right]\\
&= -y/x+x/y+(b^2-a^2)\\
&\quad \times\left[(x^2(b^2+\lambda)+y^2(a^2+\lambda))/(xy(a^2+\lambda)(b^2+\lambda))\right]\\
&= -y/x+x/y+(b^2-a^2)/xy;
\end{aligned}$$

we use (2) to make the transition from the second-to-the-last-line to the last. Thus the level curves $\lambda(x, y) = k$ satisfy

$$1/y'-y'=-y/x+x/y+(b^2-a^2)/xy, \tag{3}$$

which is the same as (1), if we set $c^2 = b^2 - a^2$. We see that as λ sweeps out the interval $[-b^2, -a^2]$ (resp. $[-a^2, \infty]$), its associated level curves are a one-parameter family of hyperbolas (respectively, ellipses) and these families are orthogonal to one another, as we would now anticipate from the invariance of (3) with respect to the transformation $y' \to -1/y'$.

We now know that the solution curves we are looking for are either hyperbolas or ellipses. A moment's reflection (if you'll excuse the pun) shows that the ellipses are extraneous solutions; light rays do not reflect off the ellipse to the focus, but

rather "refract through" at an angle that differs by 180 degrees from the reflection angle, then converge to a single focus.

Figure 5 shows that hyperbolas do have the desired reflecting property and are justly called the analogues of parabolas in the Poincaré half-plane. More succinctly (and somewhat glibly) we say that *hyperbolic parabolas are hyperbolas*. We will come back to this point in a moment.

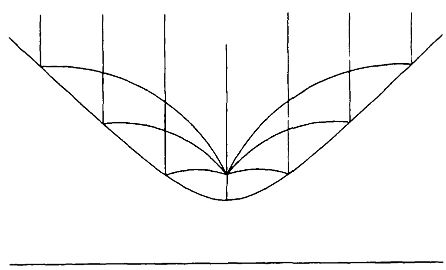


Figure 5

Figure 6 shows an interesting special case: The degenerate hyperbola $x^2 - y^2 = 0$ (consisting of two rays through the origin) reflects the vertical pencil of light rays to a single point on the boundary of the half-plane, namely the origin. Although we have not mentioned the notion of non-euclidean distance in the Poincaré half-plane, the fact is that all points along the x-axis are "infinite boundary points"; thus this degenerate "mirror" reflects parallel vertical light rays to parallel rays going out in another direction.

Are there other "non-euclidean flashlights" in hyperbolic geometry, besides the hyperbolas to which our analysis has led us? A careful rereading of our derivation should convince the reader that we have already in essence proved a uniqueness theorem: if the curve in the hyperbolic plane reflects the pencil of vertical rays onto a single point, then it must be a hyperbola. However, using some additional properties of the Poincaré half-plane that we now review, we can come up with new (and perhaps surprising) reflecting curves.

Just as the euclidean plane has rigid body motions (combinations of translations and rotations) that preserve the notion of length, so the Poincaré half-plane has analogous transformations: the linear fractional transformations that send the upper-half plane into itself. Not only do these transformations preserve the notion of distance, but they send *light ray paths to light ray paths*; classically this corresponds to the fact that linear fractional transformations send lines and/or circles to lines and/or circles. Thus all of the constructions we have been discussing in this article transform, under appropriate linear fractional transformations, to geometrically equivalent constructions in the upper-half plane. One such example is the mapping Φ: $z \to \frac{z}{z+1}$. Notice that the image of the pencil of vertical light rays under this mapping consists of a pencil of geodesic circles emanating from the point $(0, 1)$.

We apply the transformation Φ to the configuration in Figure 6 to obtain the (Poincaré equivalent) Figure 7. We can think of the light rays as emanating from $(0, 1)$, reflecting off a boundary, then going "out" to $(0, 0)$. We can also apply the transformation Φ to the configuration in Figure 5 to obtain the equivalent Figure 8. Here the reflecting boundary is the image, via Φ, of the hyperbola

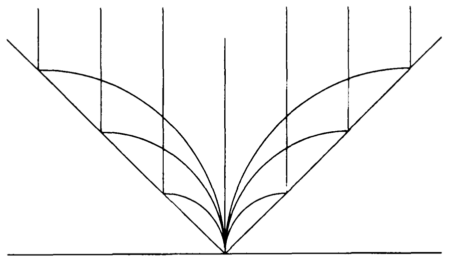


Figure 6

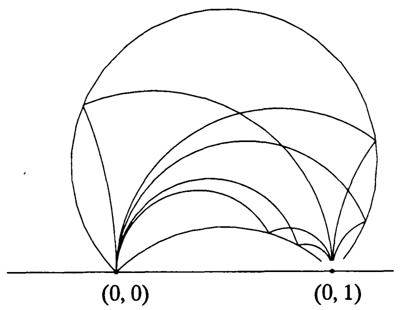


Figure 7

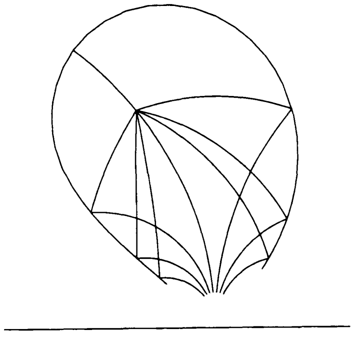


Figure 8

$x^2 - y^2 = k$ of Figure 5. It is easy to show that this image curve is algebraic (it is the level set of a quartic).

ELLIPSES AND FURTHER CONSIDERATIONS. It seems natural at this point to return to euclidean geometry and look for other geometric objects that might have analogues in hyperbolic geometry. An obvious candidate is the ellipse, with its property that light rays emanating from one focus reflect from the boundary and pass through the second focus. Given the earlier considerations of this paper, it is not too surprising that curves exist in the hyperbolic plane with an analogous two-point focussing property. It is interesting that more is true, and this topic makes contact with some quite interesting developments in classical and modern mathematics.

First we review a classical property of ellipses. Given two distinct points, we can consider the family of all ellipses (confocal) that share these two points as foci. Now imagine a billiard table built in the shape of one of these ellipses. With a piece of chalk, trace out the image of a smaller confocal ellipse on the table. Set a ball rolling on the table so that as it rolls, it just grazes (passes tangentially) the inner ellipse (Figure 9). The ball then bounces off the wall of the outer ellipse, and eventually passes by the inner ellipse again. A classical theorem on conic sections, which can be found in Salmon's treatise on the subject, implies ([S], p. 182): *each time the ball touches the inner ellipse, it passes tangentially* (*grazes*) (Salmon does not attribute this theorem to anyone, so we informally refer to it as "Salmon's theorem"). As a special case, consider the (degenerate) ellipse consisting simply of the line segment connecting the two foci. We then recover the focussing property of ellipses mentioned in the previous paragraph.

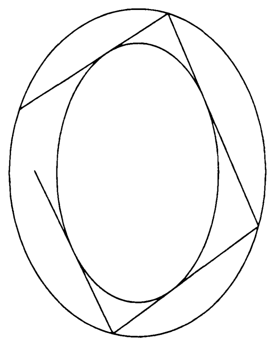


Figure 9

Salmon's theorem can also be interpreted in terms of light rays: again consider two confocal ellipses, one inner and one outer. Suppose a light ray reflects off the outer ellipse and its path just grazes the inner ellipse. Then all future crossings of the path with the inner ellipse will also be grazing.

To realize the significance of Salmon's theorem, we review the billiard problem in general. Take any (convex) curve γ, and imagine a billiard table with γ as its contour. As a ball bounces around within the interior of the curve, we need only keep track of its point of contact with the boundary, and the angle of inclination θ of the billiard ball's path, measured immediately after impact; for definiteness, we measure the angle θ between the ball's path and the interior normal ($-\pi/2 \leq \theta \leq \pi/2$). The set $M = \gamma \times [-\pi/2, \pi/2]$ is called the *phase space* and as the ball moves around the table, we obtain a sequence of points in M, which we call the *orbit* associated with the ball. Since the set M is described by two parameters (arclength s along γ and the angle θ), we say that there are *two degrees of freedom*.

It is a fact that for many curves, a typical orbit will be *dense* in M ([**K**]). But Salmon's theorem shows that *no* orbit is dense in M for an elliptical table. Why is this so? As the ball bounces inside the ellipse, we know that its behavior is constrained: it must bounce in a direction such that it grazes a fixed interior confocal ellipse. This constraint can be interpreted as saying that the orbit of the billiard ball in M must lie on a curve in M that is the graph of a function from γ to $[-\pi/2, \pi/2]$ ($\theta = \theta(\gamma(s))$), thus reducing the number of degrees of freedom from two to one.

This highly exceptional behavior of the ellipsoidal billiard table is closely connected with the theory of *completely integrable systems*. Generally, a set of differential equations is called *integrable* if its solution can be reduced, after appropriate algebraic manipulation, to antidifferentiation (also called quadrature in the classical literature). Essentially all equations encountered in an elementary differential equations course are integrable, but they are actually quite rare in practice.

There are a number of integrable systems that were known and studied in the nineteenth century, but a new and deeper understanding of these equations has been developed in the last twenty-five years, and a large number of new examples have been discovered during this time (see **[M]**, **[L]**, **[D-J]**, and the references listed there; this area of research also goes by the name of *soliton theory*).

Solutions to differential equations depend upon some continuous independent variable (say time t), but one can also introduce discrete dynamical problems, where the time parameter is an integer n, and the points $p_{-2}, p_{-1}, p_0, p_1, p_2, \dots$ are iterates of some transformation applied to the initial point p_0 (the negative indicies correspond to the inverse transformation). For example, the transformation that assigns to a point along a curve γ and a direction θ the next collision point (and reflected direction) in the billiard problem, is such a transformation.

Recently, A. Veselov has developed a theory that defines and treats *integrable transformations* or *maps*, in analogy with integrable differential equations. In a series of papers **[M-V]**, **[V1]**, **[V2]** (including joint work with J. Moser), he has discovered many discrete analogues of old and new integrable systems. His theory includes a number of geometric examples, and one result falling within the framework of his theory is that the billiard problem for the non-euclidean ellipse satisfies the analogue of Salmon's theorem: one considers two non-euclidean analogues of ellipses that are confocal, and then one can show that light rays reflecting off the outer ellipse that graze the interior ellipse at one point of contact, continue to do so for any other point of contact.

Thus, the non-euclidean flashlight has "lit the way" to one of modern mathematics' most active areas!

ACKNOWLEDGMENTS. The idea for this paper has been around for a while, and I have enjoyed discussions related to it with various colleagues. In the early stages, Troels Petersen and Marvin Greenberg were amiable listeners. Somewhat later, I had useful conversations with Joel Langer and Mark Levi during a stay at ETH, Zürich. I also thank Alexander Veselov for showing me his work, developed at that time, relating to integrable maps and billiard problems. More recently, discussions with Stan Wagon gave me the impetus to finish this up. A visit to the NSF Geometry Center at the University of Minnesota and conversations with Davide Cervone and Tamara Munzner helped bring the writing to a close. In more formal language, I am happy to acknowledge the support of the National Science Foundation Science and Technology Research Center for Computation and Visualization of Geometric Structures. Finally, I would like to dedicate this paper to my parents, "Muggy" and Lois.

REFERENCES

[B] Boothby, W., *An introduction to differentiable manifolds and Riemannian geometry*, Academic Press, New York, 1975.

[D-J] Drazin, P. G. and R. S. Johnson, *Solitons: an introduction*, Cambridge University Press, Cambridge, 1990.

[K] Kozlov, V. V., *Billiards: a genetic introduction to the dynamics of systems with impacts*, Amer. Math. Soc., Providence, R.I., 1990.

[L] Lamb, G. L., *Elements of soliton theory*, Wiley Interscience, New York, 1980.

[M] Moser, J., *Dynamical Systems C.I.M.E. Lectures*, Birkhäuser, Boston, 1980.

[M-V] Moser, J., and A. Veselov, Discrete versions of some classical integrable systems and factorization of matrix polynomials, *Comm. Math. Phys.* **139** (**2**) (1991), 217.

[S] Salmon, G., *A treatise on conic sections*, Chelsea, New York, 1954.

[S-R] Sokolnikoff, I. S., and R. Redheffer, *Mathematics of physics and modern engineering*, McGraw-Hill, New York, 1966.

[V1] Veselov, A., *What is an integrable mapping*, in *What is Integrability*?, V. E. Zakharov (Ed.), Springer-Verlag, 1990.

[V2] ______, Confocal surfaces and integrable billiards on the sphere and Lobachevsky space, *J. Geom. Phys.* **7** (**1**) (1990), 81.

Department of Mathematics & Computer Science
Drexel University
Philadelphia, PA 19104
rperline@mcs.drexel.edu

PICTURE PUZZLE

(from the collection of Paul Halmos)

A physicist who knows mathematics . . .

(see page 440)

22. ______, "Théorie des distributions et transformations de Fourier", *Ann. Univ. Grenoble* 23 (1948) pp. 7–24.
23. ______, *Théorie des distributions*. Hermann & Cie Tome I, 1950, 148 pp; Tome II, 1951, 169 pp.
24. Strichartz, Robert S., *The Way of Analysis*. Jones & Bartlett, 1995, ~ 478 pp.
25. Strauss, Walter, *Partial Differential Equations: An Introduction*. John Wiley, 1992, 425 pp.
26. Truesdell, Clifford, "The influence of elasticity on analysis: the classical heritage", *Bull. Amer. Math. Soc.* 9, (1983) pp. 293–310.
27. ______, *An Idiot's Fugitive Essays on Science*. Springer, 1984, 661 pp.
28. Walker, James S., *Fourier Analysis*. Oxford University Press, 1988, 440 pp.
29. Walker, P. L., *The Theory of Fourier Series and Integrals*. John Wiley, 1986, 192 pp.

Department of Mathematics
Indiana University Northwest
3400 Broadway
Gary, IN 46408

Answer to Picture Puzzles

(pp. 385 and 392)

George Uhlenbeck and Karen Uhlenbeck (Karen Keskula), who was once married to Olke Uhlenbeck, George's son. See the April, 1996 issue of *Math Horizons* for a profile of Karen Uhlenbeck.

On a Converse of Sharkovsky's Theorem

Saber Elaydi

In 1964, the Ukranian Mathematician Alexander Nikolaevich Sharkovsky [5] discovered a spectacular result on continuous maps on intervals. For the convenience of the reader we will state Sharkovsky's Theorem in which he used the following ordering of the set of natural numbers:

$3 \rhd 5 \rhd 7 \rhd \dots$	$2 \times 3 \rhd 2 \times 5 \rhd 2 \times 7 \rhd \dots$	$2^2 \times 3 \rhd 2^2 \times 5 \rhd 2^2 \times 7 \rhd \dots$
odd integers	$2 \times$ odd integers	$2^2 \times$ odd integers
$\dots$	$\dots$	$\dots \rhd 2^3 \rhd 2^2 \rhd 2 \rhd 1$

Here $m \rhd n$ signifies that m appears before n in the Sharkovsky ordering.

Theorem 1. (Sharkovsky [5]). *Let f: $I \to I$ be a continuous map from the interval I into itself. If $k \rhd r$ and f has a point of period k, then f must have a point of period r.*

The question that we are going to address in this note is the following: Given any positive integers k, r with $k \rhd r$, is there a continuous map that has a point of period r but no points of period k?

There are very few examples in the literature, which is scattered in many books such as [2–4]. These examples deal mostly with maps that have points of period 5 but no points of period 3 and no pattern is given to generate more examples. Moreover, examples of maps that have points of period 2^n seem missing in textbooks on dynamical systems. However, in an article by Štefan [6], a general scheme was given to generate maps that have points of period $(2n + 1)$ but no points of period $(2n - 1)$. Furthermore, using the so-called "doubling" of maps, he was able to construct maps that have points of period $2^k(2n + 1)$ but no points of period $2^k(2n - 1)$ for any positive integer n and any nonnegative integer k. Clearly, using this scheme one can generate maps that have points of period 2^k but no points of period 2^{k+1}. In this note, however, we give new and simple constructions for such maps. In addition, our proofs are very simple and should be accessible to nonspecialists. We are now ready to state our main result, which we call the Converse of Sharkovsky's Theorem.

Theorem 2. *For any positive integer r there exists a continuous map f_r: $I_r \to I_r$ on the interval I_r such that f_r has points of prime period r but no points of prime period s for all positive integers s that precede r in the Sharkovsky ordering, i.e., $s \rhd \cdots \rhd r$.*

Proof: The proof will be accomplished by the construction of the required maps. Here we have three cases to consider:

(I) odd periods,
(II) periods of the form $2^n \times$ odd natural number,
(III) periods that are powers of 2, i.e., 2^n.

Case I. Odd Periods

(a) A map that has points of period 5 but no points of period 3. Define a map f: $[1,5] \to [1,5]$ as follows:

Let $f(1) = 3$, $f(2) = 5$, $f(3) = 4$, $f(4) = 2$, $f(5) = 1$ and on each interval $[n, n+1]$ we assume f to be linear (see Figure 1).

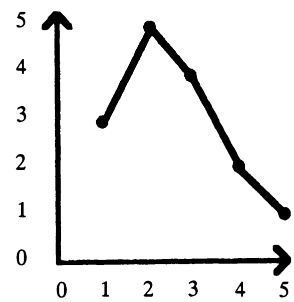


Figure 1

Observe first that none of the points 1, 2, 3, 4, 5 is a periodic point of period 3; they all belong to a 5-cycle. Notice also that

$$f^3([1,2]) = [2,5], \quad f^3([2,3]) = [3,5], \quad \text{and} \quad f^3([4,5]) = [1,4].$$

From these observations we conclude that the third iterate f^3 has no fixed points in the intervals $[1,2]$, $[2,3]$, and $[4,5]$. The situation with the interval $[3,4]$ is, however, more involved, since $f^3([3,4]) = [1,5]$. Then there are points $a, b \in [3,4]$ such that $f^3(a) = 3$, $f^3(b) = 4$. Define a map h: $[1,5] \to R$ by letting $h(x) = x - f^3(x)$. Then $h(a) \geq 0$, $h(b) \leq 0$. Hence by the intermediate value theorem, there exists a point $p \in [3,4]$ with $h(p) = 0$ or $f^3(p) = p$. We will show that p is unique and is a fixed point of f. Now $f(p) \in [2,4]$. So if $f(p) \in [2,3]$, then $f^2(p) \in [4,5]$ and thus $p = f^3(p) \in [1,2]$, which is false. Thus $f(p) \in [3,4]$ and consequently $f^2(p) \in [2,4]$. Again if $f^2(p) \in [2,3]$, then $p \doteq f^3(p) \in [4,5]$, yet another contradiction. Therefore, p, $f(p)$, and $f^2(p)$ all belong to the interval $[3,4]$. Now on the interval $[3,4]$, $f(x) = 10 - 2x$ has the unique fixed point $x^* = 10/3$. Moreover, on $[3,4]$, $f^3(x) = 30 - 8x$, which has the unique fixed point $x^* = 10/3$. Thus $p = x^* = 10/3$, and consequently f has no points of period 3.

(b) Now one can generalize this construction to manufacture continuous maps that have points of period $2n+1$ but no points of period $2n-1$ as follows:

Let f: $[1, 2n+1] \to [1, 2n+1]$ be defined by putting $f(1) = n+1$, $f(2) = 2n+1$, $f(3) = 2n$, $f(4) = 2n-1, \dots, f(n) = n+3$, $f(n+1) = n+2$, $f(n+2) = n$, $f(n+3) = n-1, \dots, f(2n+1) = 1$ (see Figure 2).

First we observe that all the integers in the interval $[1, 2n+1]$ are of period $2n+1$. For example, the orbit of the point 1 is given by the string

$$1 \xrightarrow{f} n+1 \xrightarrow{f} n \xrightarrow{f} n+2 \xrightarrow{f} n-1 \xrightarrow{f} n+3 \xrightarrow{f} n-2 \xrightarrow{f} \cdots \xrightarrow{f} 2 \xrightarrow{f} 2n+1.$$

Observe that, in addition to 1, there are two sequences of length n; one increasing: $\{n+2, n+3, \dots, 2n+1\}$; and another decreasing: $\{n+1, n, \dots, 2\}$. It remains to show that there are no points of period $2n-1$ in the interval $[1, 2n+1]$. Let us

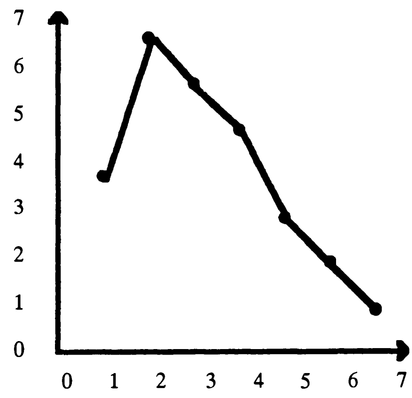


Figure 2

start with the interval [1,2]. Now $(2n-1)$ iterations of the interval [1,2] give rise to the following string:

$$[1,2] \xrightarrow{f} [n+1, 2n+1] \xrightarrow{f} [1, n+2] \xrightarrow{f} [n, 2n+1] \xrightarrow{f} [1, n+3] \xrightarrow{f}$$
$$[n-1, 2n+1] \xrightarrow{f} \cdots \xrightarrow{f} [1, 2n] \xrightarrow{f} [2, 2n+1].$$

This shows that $f^{2n-1}([1,2]) \cap [1,2) = \phi$. Hence the interval [1,2] contains no points of period $(2n-1)$. Now, we can show that all the intervals $[j, j+1]$, with the exception of the interval $[n+1, n+2]$, display the same behavior as that of the interval [1,2]. In particular, we can show that there exists an iterate of the interval $[j, j+1]$ that is precisely the interval [1,2]. Since the interval [1,2] has no points of period $(2n-1)$, it follows that the interval $[j, j+1]$ has no points of period $(2n-1)$. As for the interval $[n+1, n+2]$, notice that $f[n+1, n+2] = [n, n+2]$. Hence there are two cases for $x \in [n+1, n+2]$.

Case (a) $f^k(x) \in [n+1, n+2]$ for all $k \in Z^+$. Since $|f'| > 1$ on the interval $[n+1, n+2]$, it follows that x is actually a fixed point of f.

Case (b) $f^k(x) \notin [n+1, n+2]$ for some positive integer k. Then $f^k(x) \in [n, n+1]$ and by the previous analysis an iterate of x lies in the interval [1,2].

In either case, x cannot have period $(2n-1)$.

Case II. Maps that have points of period $2^k(2n+1)$ but no points of period $2^k(2n-1)$.

Let us start with period 2×5 but not 2×3. We consider first the map $f\colon I \to I$, $I = [1,5]$ which was considered in Case Ia (Figure 1).

Define a new map $g\colon [1,13] \to [1,13]$ as follows:

$$g(x) = \begin{cases} f(x) + 8, & \text{for } 1 \le x \le 5 \\ x - 8, & \text{for } 9 \le x \le 13 \end{cases}$$

and for $5 < x < 9$, we connect the points (5,9) and (9,1) by a line (see Figure 3).

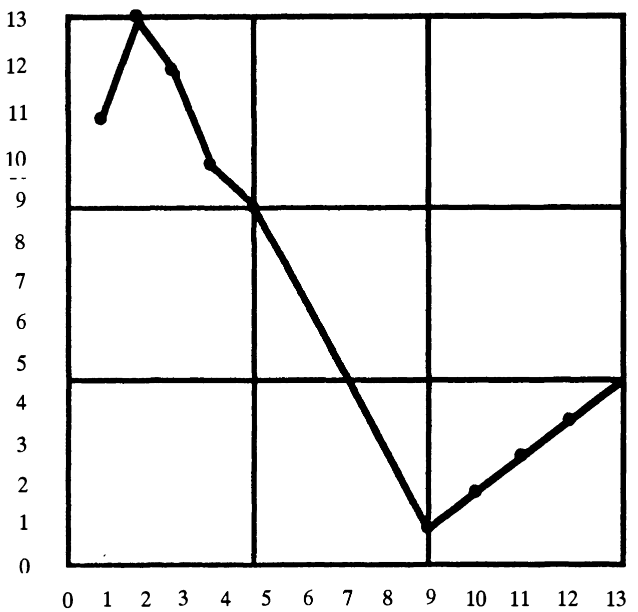


Figure 3

The map g is called the "double of f." Observe first that none of the points $1, 2, 3, 4, 5, 9, 10, 11, 12, 13$ is of period 6; they all belong to a 10-cycle. Moreover, if $x \in [1, 5]$, then $g(x) \in [9, 13]$ and $g^2(x) = f(x)$. Since f has no points of period 3, it follows that g has no points of period 6 in the interval $[1, 5]$. Since $g[9, 13] = [1, 5]$, it follows that no point in the interval $[9, 13]$ is of period 6. The situation with the interval $[5, 9]$ requires a different argument. Since $g^6[5, 9] = [4, 10]$, it follows by an argument similar to that used in Case I(a) that g^6 has a fixed point $p \in (5, 9)$. Now for any n, $1 \leq n \leq 5$, $g^n(p) \notin (5, 9)$, then $g^{n+r}(p) \in [1, 5] \cup [9, 13]$ for all $r > 0$. This implies that $g^6(p) \neq p$, a contradiction. Thus $p, g(p), \ldots, g^5(p) \in (5, 9)$. By simple computations, one can show that the only fixed point of $g, g^2, \ldots, g^6$ is $p = 19/3$. Thus, g has no points of period 6.

The general procedure for constructing a map that has points of period $2(2n + 1)$ but no points of period $2(2n - 1)$, $n = 1, 2, 3, \ldots$ may be explained as follows. We start with a map f: $[1, 1 + h] \to [1, 1 + h]$ with points of period $(2n + 1)$ but no points of period $(2n - 1)$. We define the double map g: $[1, 1 + 3h] \to [1, 1 + 3h]$ as follows:

$$g(x) = \begin{cases} f(x) + 2h, & \text{for } 1 \leq x \leq 1 + h \\ x - 2h, & \text{for } 1 + 2h \leq x \leq 1 + 3h \end{cases}$$

and by linearity for $1 + h < x < 1 + 2h$. Repetition of the preceding scheme would create maps with points of period $2^k(2n + 1)$ but no points of period $2^k(2n - 1)$, $k = 2, 3, 4, \ldots$

Case III. Periods of the form 2^n

(a) A map that has points of period 2 but no points of period 2^2. Let f: $[1, 2] \to [1, 2]$ be defined by $f(x) = -x + 3$. Here every point, except the fixed point $3/2$, in the interval $[1, 2]$ is of prime period 2. Hence there are no points in the interval $[1, 2]$ with prime period 2^2.

(b) A map that has points of period 2^2 but no points of period 2^3. Let $f\colon [1,4] \to [1,4]$ be defined as follows: $f(1) = 3$, $f(2) = 4$, $f(3) = 2$, $f(4) = 1$ and on each interval $[n, n+1]$ we assume f to be linear (Figure 4).

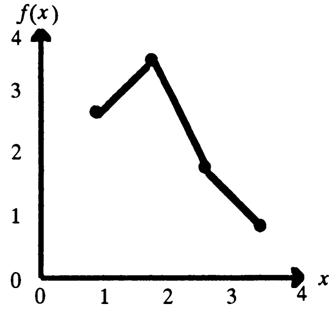


Figure 4

Notice that $f([1,2]) = [3,4]$ and $f([3,4]) = [1,2]$, and f is linear on $[1,2]$ and $[3,4]$. Thus $f^2([1,2]) = [1,2]$ and $f^2([3,4]) = [3,4]$. Also, f^2 is decreasing, as f^4 is increasing. Therefore $f^4(x) = x$ for all $x \in [1,2] \cup [3,4]$.

Hence every point in the interval $[1,2]$ is of prime period 4 except the point $3/2$, which is of prime period 2. Similarly, every point in the interval $[3,4]$ is of prime period 4 except the point $7/2$, which is of prime period 2. Next we deal with the interval $[2,3]$. Since $f[2,3] = [2,4]$, points in the interval $[2,3]$ either leave the interval $[2,3]$ after many iterations or stay in the interval $[2,3]$ for all iterations. Now if for a point $x \in [2,3]$, and for some $k \in Z^+$, $f^k(x) \in [1,2] \cup [3,4]$, then its orbit will be attracted to either a 4-cycle or a 2-cycle. On the other hand if the orbit of $x \in [2,3]$ is a subset of the interval $[2,3]$, then $f^n(x) = f_2^n(x)$, where $f_2(x) = -2x + 8$. But $f_2^8(x) = 256x - 680$ has the fixed point $x^* = 8/3$, which is a fixed point of the map f. Hence the map f has no points of period 8 or any other periods that precede it in the Sharkovsky order.

(c) To construct maps that have points of period 2^n but no points of period 2^{n+1}, we use the double map g that was used previously in Case II. Here we start with the map f defined in Case IIIb, which has points of period 2^2 but no points of period 2^3. The double map $g\colon [1,10] \to [1,10]$ is defined as follows:

$$g(x) = \begin{cases} f(x) + 6, & \text{for } 1 \le x \le 4 \\ x - 6, & \text{for } 7 \le x \le 10. \end{cases}$$

Then the map g has points of period 2^3 but no points of period 2^4 (Figure 5).

This construction can be carried out indefinitely to produce maps that have points of period 2^n but no points of period 2^{n+1}

Remark. There is still one more question to be settled. Can we construct maps that have points of period $2^n \times 3$ but no points of any period of the form $2^{n-1} \times$ odd integer? Fortunately, using the double map one can give an affirmative answer to this question. Let us first construct a map that has points of period 2×3 but no points of odd periods. Define $f\colon [1,3] \to [1,3]$ by letting $f(1) = 2$, $f(2) = 3$, and $f(3) = 1$ and on each interval $[n, n+1]$ we assume f to be

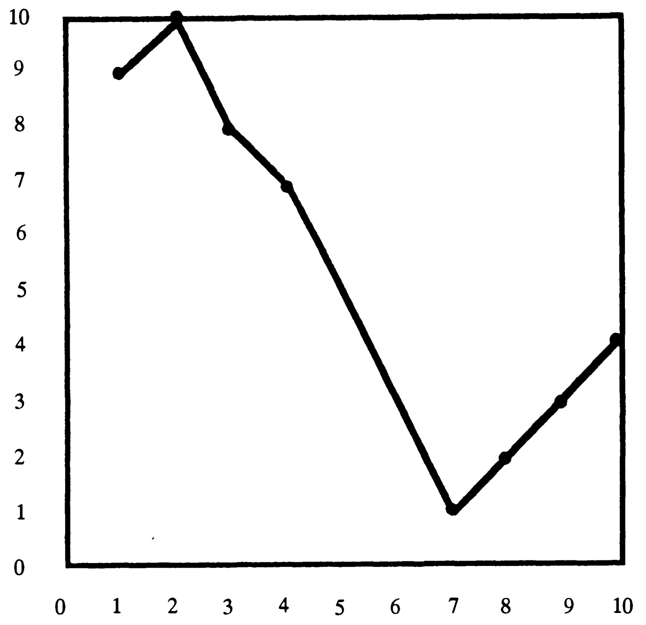


Figure 5

linear. Clearly the points 1, 2, and 3 are all of period 3. Now the double map g: $[1, 7] \to [1, 7]$ is defined as

$$g(x) = \begin{cases} f(x) + 4, & \text{for } 1 \leq x \leq 3 \\ x - 4, & \text{for } 5 \leq x \leq 7. \end{cases}$$

Observe that the map g has points of period 2×3 but no points of odd period. By repeating this process, one can construct continuous maps that have points of period $2^n \times 3$ but no points of period $2^{n-1} \times$ odd integer.

Addendum. After writing this note, I was informed by Dr. Hasfura of Trinity University that Delahaye [1] had an example of a continuous map that has points of period 2^n for all nonnegative integers n and no other periods. For the sake of completion, I include this example here.

Example. Let $I = [0, 1]$ and $I_k = [1 - 1/3^k, 1 - 2/3^{k+1}]$, for all $k \geq 0$. For each k let f_k: $I_k \to I_k$ be a continuous map. Define a continuous map f: $I \to I$ by letting $f(1) = 1$, $f(x) = f_k(x)$ if $x \in I_k$ and by linearity elsewhere. Now for each nonnegative integer k choose f_k such that it has points of period 2^k but no points of period 2^{k+1}. Then f has points of periods 2^n for all nonnegative integers n but no points of other periods.

ACKNOWLEDGMENTS. I would like to thank Professor Dr. Bernd Aulbach for raising the question that led to writing this note, during my visit to the Institut für Mathematik der Universität Augsburg. My special thanks go to Professor Denny Gulick who read thoroughly this manuscript and made many suggestions that improved it in both style and content.

REFERENCES

1. Delahaye, J. P., Fonctions admittant des Cycles d'ordre n'importe quelle puissance de 2 et aucun autre cycle, *C.R. Acad. Sci. Paris*, Sér. A–B, 291 (1980) A 323–325.
2. Devaney, R., *An introduction to chaotic dynamical systems*, Second Edition, Addison-Wesley, Menlo Park, California, 1989.
3. Elaydi, S., *Chaotic dynamical systems with fractals*, Preprint.
4. Gulick, D., *Encounters with chaos*, McGraw-Hill, New York, 1992.
5. Sharkovsky, A. N. Co-existence of Cycles of a continuous mapping of a line into itself. *Ukranian Math. Z.* 16 (1964), 61–71.
6. Štefan, P., A Theorem of Sharkovsky on the existence of periodic orbits of continuous endomorphisms of the real line, *Comm. Math. Phys.* 54 (1977), 237–248.

Department of Mathematics
Trinity University
San Antonio, TX 78212
selaydi@trinity.edu

PICTURE PUZZLE
(from the collection of Paul Halmos)

. . . and a mathematician who knows physics.
(see page 440)

22. ______, "Théorie des distributions et transformations de Fourier", *Ann. Univ. Grenoble* 23 (1948) pp. 7–24.
23. ______, *Théorie des distributions*. Hermann & Cie Tome I, 1950, 148 pp; Tome II, 1951, 169 pp.
24. Strichartz, Robert S., *The Way of Analysis*. Jones & Bartlett, 1995, ~ 478 pp.
25. Strauss, Walter, *Partial Differential Equations: An Introduction*. John Wiley, 1992, 425 pp.
26. Truesdell, Clifford, "The influence of elasticity on analysis: the classical heritage", *Bull. Amer. Math. Soc.* 9, (1983) pp. 293–310.
27. ______, *An Idiot's Fugitive Essays on Science*. Springer, 1984, 661 pp.
28. Walker, James S., *Fourier Analysis*. Oxford University Press, 1988, 440 pp.
29. Walker, P. L., *The Theory of Fourier Series and Integrals*. John Wiley, 1986, 192 pp.

Department of Mathematics
Indiana University Northwest
3400 Broadway
Gary, IN 46408

Answer to Picture Puzzles

(pp. 385 and 392)

George Uhlenbeck and Karen Uhlenbeck (Karen Keskula), who was once married to Olke Uhlenbeck, George's son. See the April, 1996 issue of *Math Horizons* for a profile of Karen Uhlenbeck.

Splitting a Polygon into Two Congruent Pieces

Kimmo Eriksson

1. INTRODUCTION. I have run into the following kind of problem in an entertainment problem column: Show how to split the polygon in Fig. 1(a) into two congruent pieces!

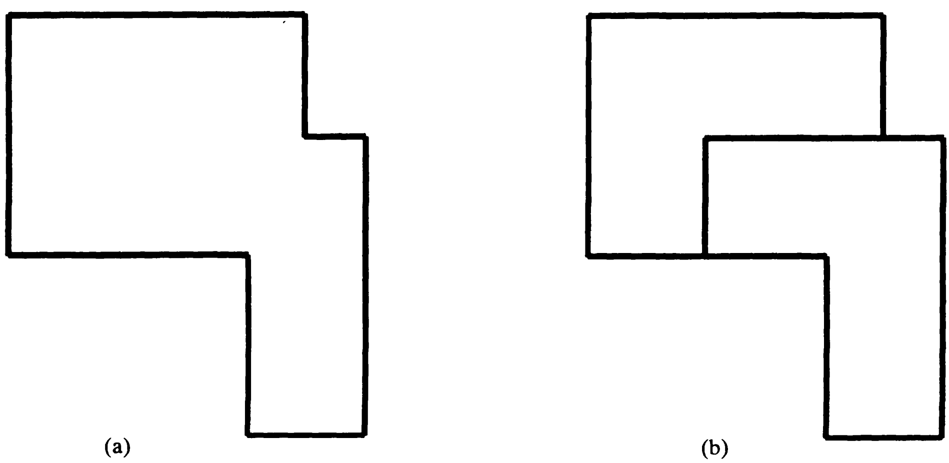


Figure 1. (a) Problem: Split the polygon in two congruent pieces! (b) Solution.

It is quite easy to find the solution to this particular splitting problem (as shown in Fig. 1(b)), but there are certainly cases where the solution is far from obvious; at a graduation party in Stockholm 1994 (with a lot of mathematicians among the guests), nobody was able to solve all the problems in the column mentioned above. On such occasions an effective algorithm that solves the problem would come in handy. It is the aim of this paper to provide such an algorithm. At the end of the paper I show that a theorem on splitting convex shapes follows from the algorithm. I also give a few nice problems for the reader to practise on.

According to my friend Doron Zeilberger, this kind of problem goes back to the great puzzle inventor Henry Dudeney who wrote several books with mathematical amusement problems in the first decades of this century, e.g. [**1**]. In a recreational math book by Fred Schuh [3], I found a puzzle of this kind, and also one variant where the polygon should be split in *four* congruent quarters. I do not claim to have an algorithm for variants with more than two parts, but I would love to see one! After writing the first version of this paper, I observed an instance of the splitting problem on one of a series of mathematical posters on the wall by the MIT math department staircase. This poster referred to an eighteen year old

Mathematical Games column in Scientific American, where Martin Gardner discusses the splitting problem and gives a dozen problems for the readers. Gardner writes: "As far as I know there is no algorithm for deciding in general whether a shape can be divided into two or more congruent parts, and interesting theorems about such divisions are curiously scarce."

But let us return to the party. There were arguments among the mathematicians present concerning the meaning of the word "congruence". Some people hold to the meaning "equivalence up to translations and rotations", which we will here call *proper congruence*, and some are more open-minded and also include mirror reflections. We will call objects *mirror congruent* if they are congruent in the wide meaning but are not properly congruent. The same algorithm can be used to find both properly congruent pieces and mirror congruent pieces, with a minor adjustment for the latter case, which will be enclosed in brackets: [...]. The proofs below will contain the details only for the proper case, and just hint on how the mirror case is handled. Indeed, the same algorithm can be used also for "pseudo-polygons" where the border segments are not straight but curved. See a remark at the end. Fig. 7 shows a pseudo-polygon with a mirror congruent splitting. In fact, the algorithm also works for polygons with *holes*, as in Fig. 8(b), but let us not burden the proof with these generalities.

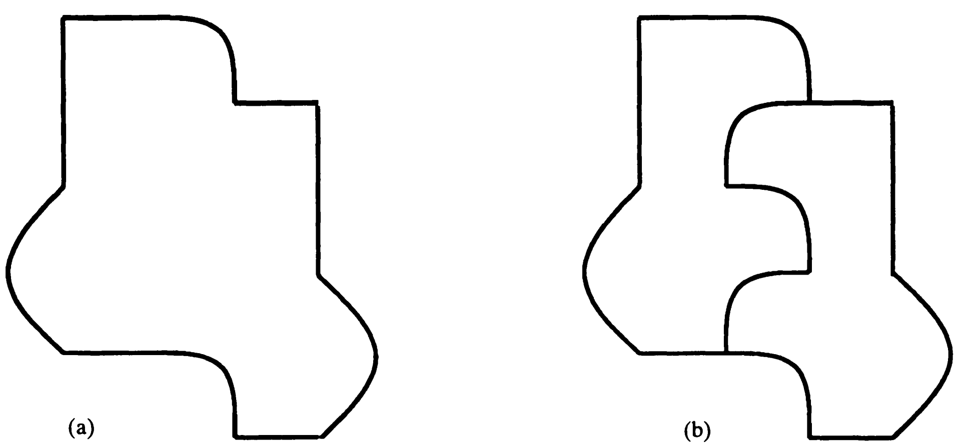


Figure 2. A pseudo-polygon splittable in mirror congruent pieces.

As the party discussion showed, we need to be precise about our definitions, in particular about non-specialist words. So, by a *polygon* we mean a connected area in the plane (without holes) whose border is a finite collection of line segments. An intersection point between two such line segments is an *endpoint*. At every endpoint the polygon has an angle in the open interval $]0, 360°[$. An endpoint where the angle is not 180° is a *corner*. The *midpoint* of a line segment is the point equally far from the two end points.

In the algorithm we will be dealing with paths. To make the terminology fit the picture, we say that two paths in the plane travel *in parallel* if they are congruent (properly or mirror).

2. THE ALGORITHM. Algorithm. We shall determine if a given polygon P can be split in two properly congruent [or mirror congruent] connected pieces. Let S

be the set of points consisting of all endpoints of the bordering line segments of P, together with the midpoints of those line segments whose two angles sum up to 180°. For every pair $\{p_1, p_2\}$ of points in S we shall try to split P by constructing two paths, starting in p_1 and p_2, that travel in parallel. We make two tries, one where both paths travel clockwise and one where both go anticlockwise. [And for mirror congruence we make two more tries, where the paths go in opposite directions.] Let Path 1 start running in the selected direction along the border from p_1, and similarly for Path 2 starting from p_2. Path 1 and Path 2 are not allowed to intersect in more than two points along the border of P. Thus the paths cannot travel in parallel forever along the border. Let them do so as long as possible, that is, until either (1) one of the paths, say Path 1, must turn away from the border into the interior of the polygon for the other one to be able to stay on the border or else (2) Path 1, say, reaches p_2.

If in case (2) but not in case (1), then if Path 2 has not at the same time reached p_1, we give up this try; otherwise we have 180° rotational symmetry and get the desired splitting by a straight cut from p_1 to p_2 (or, if this straight line is not contained in the interior of P, use only the center portion of it). [When the paths go in different directions, case (2) is when the paths meet at a point p_3. If so, then see if it is possible to continue the paths in parallel until they meet again in a second point p_4. If so, there is mirror symmetry; split by a straight cut through p_3 and p_4. Otherwise, give up.]

In case (1), let Path 2 continue along the border with Path 1 travelling in parallel in the interior until Path 1 reaches the border once again. [In the mirror congruent case Path 2 might meet Path 1 in which case it shall proceed along Path 1 rather than along the border.] Path 1 has now cut through the polygon, and this cut is our candidate for splitting P in congruent pieces. It is easy to check if it is a successful candidate: just check if it is possible to continue the paths travelling in parallel such that Path 1 ends up in p_1 enclosing one piece, while Path 2 (following the cut made by Path 1) ends up in p_2 enclosing the other piece.

Remark. To analyze the complexity, let n be the number of end points. (The input to the algorithm is the polygon, that is, the endpoints.) Since the number of midpoints considered at most equals the number of end points, the total number of pairs to consider is $O(n^2)$. For every pair, the time for the algorithm to construct the paths is clearly proportional to the number of end points, so the total complexity is $O(n^3)$.

Fig. 3 shows two examples of how the algorithm works for different choices of starting points in the same polygon.

Theorem 1. *The algorithm solves the splitting problem, i.e. the polygon P is splittable in two congruent pieces if and only if the algorithm finds a successful cut.*

The "if" part is a trivial statement. We will prove the "only if" part of the theorem via a couple of lemmas, using no more than completely elementary geometry and common sense. It is convenient to have the following definitions.

Definition. If P is splittable in two congruent pieces P_1 and P_2, we say that a pair of subsets $S_1 \subset P_1$, $S_2 \subset P_2$ are *split-congruent* if the sets S_1 and S_2 coincide when the pieces are identified (i.e. placed one on top of the other with a perfect fit).

By the *P-border* of a piece P_i, we mean the intersection of the border of P with P_i.

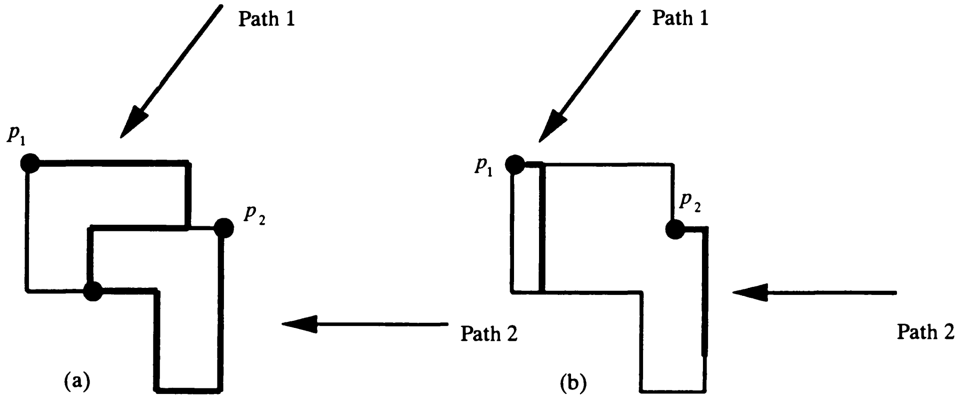


Figure 3. Running the algorithm with two different pairs of points p_1, p_2: (a) A successful cut. (b) An unsuccessful cut.

Lemma 2. *Suppose P can be split successfully in P_1 and P_2. Then there exists a pair of split-congruent segments (of length greater than zero) of the P-borders of P_1 and P_2.*

Proof: What would a counterexample look like? A polygon P, split in two congruent pieces P_1 and P_2, where the P-border A of P_1 is such that the segment A' of the border of P_2 that is split-congruent to A has no part in common with the border of P but is instead entirely part of the cut. Thus, the border of P_1 consists of segments A and A', possibly connected by some segments B and C that by the definition of A must also be part of the cut. See Fig. 4(a). The pieces are congruent, so the border of P_2 consists of segments A', A'', B' and C' where the apostrophe signifies congruence. See Fig. 4(b). But by assumption, when the pieces are connected, segments B and C shall have no part in common with the border of P! This is possible only if B and C are empty, yielding a forbidden degenerate polygon containing no area, see Fig. 4(c).

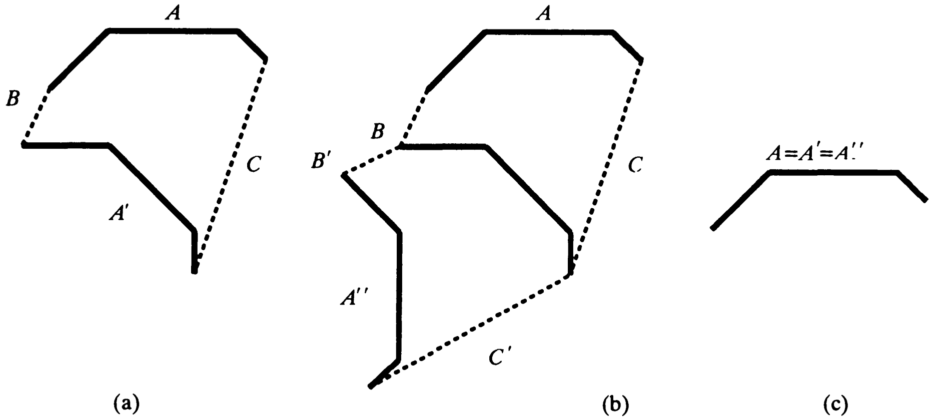


Figure 4. A sketch of the failed counterexample: (a) P_1. (b) P split in congruent pieces P_1 and P_2. (c) A forbidden degenerate polygon.

[In the mirror congruent case, we instead have A' mirror congruent to A and A'', and B connected to C' and C connected to B'. Again, it is absurd that B and C have no part in common with the border of P.] ■

We are looking for a pair of split-congruent starting points for the algorithm, and Lemma 2 has provided us with two split-congruent border segments for the next lemma to search in. A point in a segment that is not an endpoint of the segment is called an *inner point* below.

Lemma 3. *Suppose P can be split successfully. Then there exists in the set S a pair of split-congruent points. At least one of these points is an inner point of the P-border of its piece.*

Proof: Let P_1 and P_2 be the two congruent pieces. Lemma 2 guarantees the existence of maximal split-congruent segments $\overline{pq}$ and $\overline{p'q'}$ of the border of P, lying in P_1 and P_2 respectively. If the segment $\overline{pq}$ contains a corner v in its interior, then there is a split-congruent corner v' in $\overline{p'q'}$ and we are done, since all corners are endpoints, and all endpoints belong to S. Suppose in the following that there are no such corners, so $\overline{pq}$ and $\overline{p'q'}$ are line segments.

As usual we treat the properly congruent case first: The segments were chosen to be maximal, so one of them, say $\overline{pq}$, must be adjacent to the cut on the clockwise side, i.e. at q, and the other segment $\overline{p'q'}$ must be adjacent to the cut on the anticlockwise side, i.e. at p'. Hence p' is the same point as q, an intersection point of the border of P with the cut. Since $\overline{pq}$ and $\overline{p'q'}$ are line segments, and the polygons are nondegenerate, p is an inner point of the P-border of P_1, and q is likewise an inner point of the P-border of P_2. This implies that p is a corner of P_1 if and only if it is a corner of P, and analogously for q'.

We shall now show that either both p and p' belong to S, or both q and q' belong to S. We separate two cases:

Case 1: *The point $p' = q$ is a corner of P.* We claim that either p or q' must be a corner of P. Suppose p is not a corner of P. Then p is not a corner of P_1 either, so by congruence p' is not a corner of P_2. See Fig. 5(a). But then, since the angle of P at $p' = q$ is less than 360°, q must be a corner of P_1. By congruence, q' is a corner of P_2 and hence a corner of P.

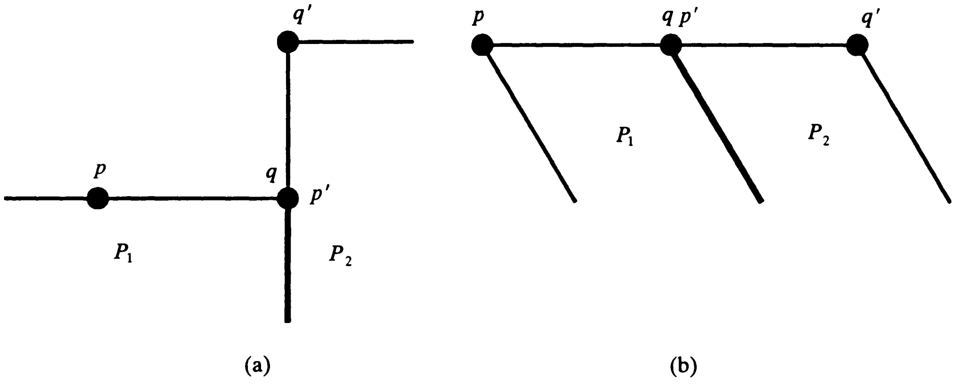


Figure 5. Line segments $\overline{pq}$ and $\overline{p'q'}$ where $p' = q$ is either (a) a corner, or (b) not a corner.

Case 2: *The point $p' = q$ is not a corner of P.* Then the cut makes q a corner of P_1 and p' a corner of P_2, see Fig. 5(b). But since P_1 and P_2 are congruent, this means that p is a corner of P_1 and hence of P, and q' is a corner of P_2 and hence of P. But then $p' = q$ is the midpoint of the line segment of the border of P between the corners p and q', and the sum of the angles at p and q' must be 180°, so by definition of the set S we have that $p' = q$ also belongs to S.

[In the mirror congruent case, p' and q are the two different points where the cut intersects the border. If, say, p' is a corner, and p is not, then the argument is a variation of the one above where instead of q' we consider the point p'' in P_2 split-congruent to p' in P_1. If neither p' nor q is a corner, and neither $\overline{pq}$ nor $\overline{qp''}$ contain any inner corners, then it is not hard to show that the polygon must look like in Fig. 6, with two parallel line segments A and A', connected by two parallel segments B and B', and the algorithm will in fact split it in two properly congruent pieces instead of two mirror congruent pieces.] ■

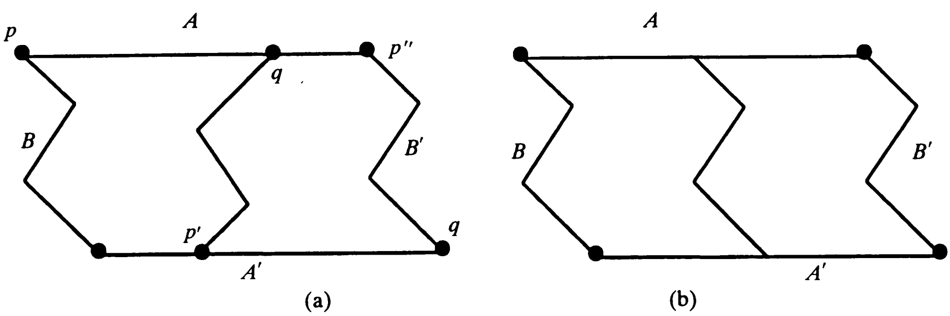


Figure 6. A polygon that is splittable both in (a) mirror-congruent pieces, and (b) ordinarily congruent pieces.

Proof of Theorem 1: Thanks to Lemma 3, we know that a polygon splittable in congruent pieces P_1 and P_2 will have a pair of split-congruent points in S, $p_1 \in P_1$ and $p_2 \in P_2$, with, say, p_1 in the interior of the P-border of P_1. From the point p_2, in at least one direction of the border of P_2 one follows the P-border of P_2, say clockwise. Theorem 1 follows immediately: The algorithm will try travelling in parallel clockwise from p_1 and [anticlockwise] from p_2, and since these points are split-congruent, such paths will automatically follow the borders of P_1 and P_2. With the exception of the case when both paths must dive into the interior at the same time, which is the case of rotational symmetry [mirror symmetry]. ■

Remark. As promised in the beginning, let's discuss what happens when we deal with pseudo-polygons, whose borders consist of segments (we want at least two) that are not necessarily straight, but at least differentiable. Well, the proof of Lemma 2 does not rely on the line shape of the segments. And since midpoints and angles are well-defined also for well-behaved shapes other than line segments, the proof of Lemma 3 does rely on the shape for one issue only: that p and p' cannot both be cut points. This could in fact be the case for other segment shapes, if the border just consists of two congruent curved segments so that the area has 180° rotational symmetry in the properly congruent case, and mirror symmetry in the mirror congruent case. But these cases are already treated by the algorithm, so everything works!

3. A COROLLARY AND SOME EXERCISES. Let us go back to the algorithm and the simpler case of proper congruence. In the properly congruent nonsymmetric case it is worth noting that Path 1 and Path 2 never at the same time pass through the interior of P. Hence, if in this case the algorithm succeeds in splitting P into properly congruent pieces P_1 and P_2, then the segment of the border of P_1 that coincides with the cut is split-congruent with a segment of the P-border of P_2, and vice versa. This leads us to the following concrete result.

Corollary 4. *A convex polygon is splittable in two properly congruent pieces if and only if it has rotational symmetry.*

Proof: A polygon has rotational symmetry if and only if it is splittable in two properly congruent pieces by a straight cut. So, suppose that the algorithm has succeeded in splitting a convex polygon P into properly congruent pieces P_1 and P_2 with a nonstraight cut. We must show that P is not convex. Since the cut is not straight it has some corner q, which must then be a concave corner of one of the pieces, say P_1. But from the observation above we know that this concave corner must be split-congruent to a concave corner of the P-border of P_2, and consequently P is not convex. ∎

However, when mirror congruence is allowed it is no longer true that the polygon must be symmetric. A counterexample is shown in Fig. 7.

Exercises. In Fig. 8 are shown three splittable polygons for you, the reader, to practise on—either your problem solving capabilities or your patience to run the algorithm. Both properly congruent and mirror congruent splits are represented among these three problems. Observe that the polygon (b), though containing a connected area, has a *hole*. The algorithm works for polygons with holes too!

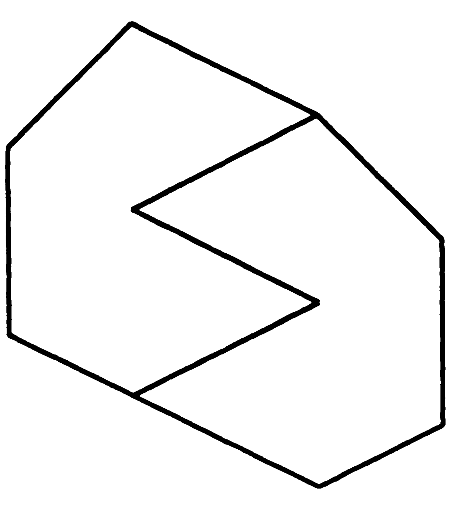


Figure 7

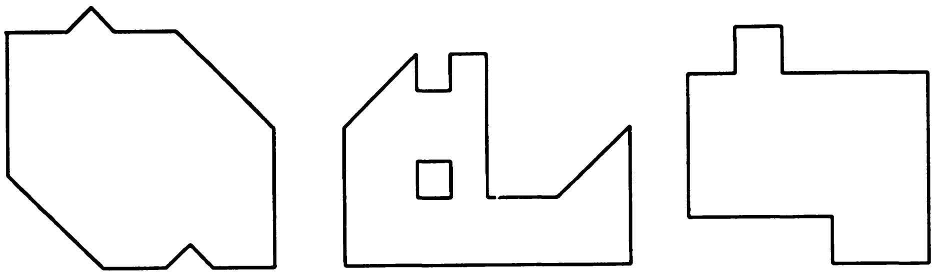


Figure 8. Three polygon splitting problems—mirror congruence allowed!

ACKNOWLEDGMENTS. I am grateful to the anonymous referee for helpful suggestions, and in particular for conjecturing the corollary above. I thank Henrik Eriksson, Jonas Sjöstrand, and Doron Zeilberger for valuable comments.

REFERENCES

1. Dudeney, H. E., *Amusements in Mathematics*, 1917, reprinted by Dover publications, 1957.
2. Gardner, M., Mathematical Games, *Scientific American* **237** (1977), 132–137.
3. Schuh, F., *The Master Book of Mathematical Recreations*, Dover publications, 1968. Puzzle 276.

Department of Mathematics
Stockholm University
S-106 91 Stockholm
SWEDEN
kimmo@nada.kth.se

I was very quick at math. I could quickly grasp any problem they threw at me and solve it in my head. I often stood in for Lydia Mikhailovna when she had to go into the city or run her own errands. She trusted me to conduct the math class. I taught basic arithmetic to the other children.

N. S. Khrushchev, *Khrushchev Remembers: The Glasnost Tapes*
Little, Brown and Co., Boston, 1990, pp. 5–6.

Contributed by Dominico Rosa, Teikyo Post University

A Geometric Approach to Determinants

John Hannah

We are all happy to use pictures when we first introduce students to calculus. Why not take the same approach in linear algebra? While good progress has been made in this direction in recent years, determinants seem to have escaped this trend. Most textbooks still introduce them via cofactor expansions (see **[FB]** and **[N]** for example), the permutation definition (**[AK]**, **[Se]**), or via their alternating multilinear form properties (**[DL]**, **[St]**).

In this article I propose a geometric introduction to determinants. The details are not new, though they are well scattered through the literature. For example, a geometric view of the 2×2 case is used as motivation for an algebraic approach in **[DL]** and **[O]**. What perhaps is new (or at any rate, has not been fashionable for at least a couple of generations) is that I am suggesting that the geometric view be given a defining role similar to that given the "area under the graph" definition of the integral, which we routinely use in beginning calculus courses.

Before you push the panic button, I'm not suggesting that rigorous algebraic approaches be abandoned. What I am saying is that, particularly in the case of determinants, this approach is not very suitable for students who are meeting linear algebra for the first time. In fact, many textbooks implicitly recognize this problem by relegating some key proofs to later sections or appendices, so that students may avoid them or, at any rate, take them on trust (see **[FB]**, **[N]**). In my own institution, I see a geometric approach as being appropriate for our first year linear algebra students, while an algebraic approach is more appropriate for our advanced courses.

Just as in the calculus context, geometry helps students to form mental images or constructs that they can use to help them understand what determinants are all about. I have anecdotal evidence that it encourages students to engage in what Blum and Kirsch call "preformal" proving **[BK]**. In other words, students can see or conjecture properties of the determinant, along with (geometric) explanations appropriate to their level of mathematical development.

Another reason for this approach is that I want a treatment that focuses on the important properties of the determinant, without getting involved in issues that I see as being peripheral for first year students, most of whom will not major in mathematics. Despite this, I like to think that my shopping list will keep most mathematicians happy, too. Here is my list of "what every first year student should know about determinants"

1. $\det A$ gives the area or volume magnification factor for the linear transformation $\mathbf{x} \to A\mathbf{x}$;
2. A is invertible if and only if $\det A \neq 0$;
3. $\det(AB) = (\det A)(\det B)$;
4. the most efficient way of calculating $\det A$ is by row (or column) reducing A to triangular form.

In what follows I have sketched out the main features of this approach to determinants. I have tried to retain some of the flavour of the way I present it to students, but of course I can leave out some of the details when I am talking to you.

DEFINING THE DETERMINANT. Let A be a $n \times n$ real matrix. We can view A as a linear transformation from R^n to R^n given by $\mathbf{x} \to A\mathbf{x}$.

If A is a 2×2 matrix with column vectors $\mathbf{a}$ and $\mathbf{b}$, then the linearity means that A transforms the unit square in R^2 into the parallelogram in R^2 determined by $\mathbf{a}$ and $\mathbf{b}$ (see Figure 1).

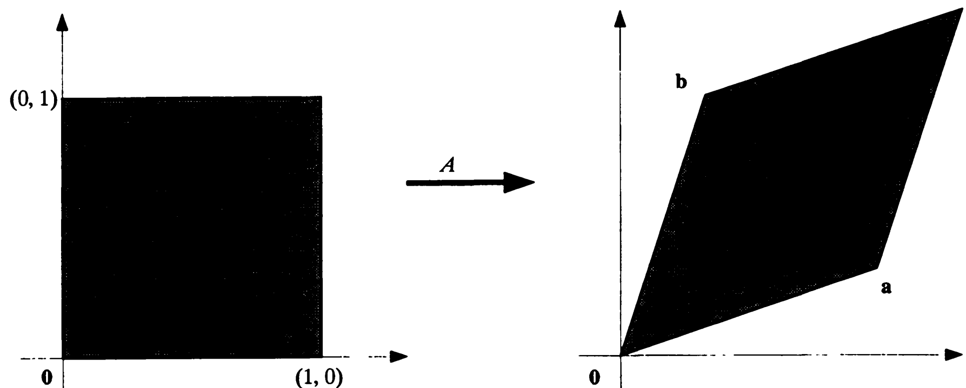


Figure 1. Effect of the matrix $A = (\mathbf{a}|\mathbf{b})$ on the unit square.

Similarly in the 3×3 case, A maps the unit cube in R^3 into the parallelepiped (or solid parallelogram) in R^3 determined by the column vectors of A. In general, an $n \times n$ matrix A maps the unit n-cube in R^n into the n-dimensional parallelogram determined by the column vectors of A.

Other squares (or cubes, or hypercubes, etc.) are transformed in much the same way and scaling the sides of the squares merely scales the sides of the parallelograms (or parallelepipeds, or higher dimensional parallelograms) by the same amount. In particular, the magnification factor

$$\frac{\text{area (or volume) of image region}}{\text{area (or volume) of original region}}$$

is always the same, no matter which squares (or cubes, or hypercubes) we start with. Indeed, since we can calculate the areas (or volumes) of reasonably nice regions by covering them with little squares (or cubes) and taking limits, the above ratio will still be the same for these regions, too.

Definition. The determinant of the matrix A is the above magnification factor.

For example, since the unit square has area 1, the determinant of a 2×2 matrix A is the area of the parallelogram determined by the columns of A. Similarly, the determinant of a 3×3 matrix A is the volume of the parallelepiped determined by the columns of A. Notice that this way of defining the determinant does not give us an obvious way of calculating its value!

Convention: In what follows I shall use *cube*, *volume* and *solid parallelogram* as a sort of dimension-free shorthand for the corresponding n-dimensional concepts.

DETERMINANTS AND MATRIX MULTIPLICATION. From the transformation point of view, matrix multiplication corresponds to function composition. If we use this to calculate the magnification factor for AB, we get the product rule

$$\det(AB) = (\det A)(\det B).$$

Now suppose that A is an invertible $n \times n$ matrix. Since the transformation corresponding to the $n \times n$ identity matrix I clearly leaves everything where it is, we have $\det I = 1$. Hence the product rule shows that $(\det A^{-1})(\det A) = 1$. So $\det A$ must be nonzero and

$$\det(A^{-1}) = \frac{1}{\det A}.$$

What if A is not invertible? Since the reduced row echelon form of A is not the identity matrix, the system $A\mathbf{x} = \mathbf{0}$ has an infinite number of solutions, so there must be some nonzero $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$. Now imagine A transforming a cube having one of its sides parallel to $\mathbf{x}$. Since one of the edges of the image solid parallelogram has zero length, its volume must be zero. Hence $\det A = 0$. This means that the determinant can be used to test the invertibility of a matrix:

$$A \text{ is invertible if and only if } \det A \neq 0.$$

AN UNEXPECTED DEVELOPMENT. We want a practical way of evaluating $\det A$. We are going to use column operations (and later, row operations) to simplify the matrix. The idea here is that column operations on the matrix A correspond to geometric operations on the solid parallelogram that we have used to define $\det A$. Hence we can calculate the effect of these operations on the appropriate volumes.

We begin with the most innocuous-looking of the elementary column operations.

Swapping two columns: It is tempting to think that this operation does not change the image, and so the determinant is not affected. However, it turns out that we have to choose between this idea and the equally attractive idea that $\det A$ is an additive function if we let just one column vary at a time. For example, letting just the third column vary would give

$$\det(\mathbf{a}|\mathbf{b}|\mathbf{c} + \mathbf{d}) = \det(\mathbf{a}|\mathbf{b}|\mathbf{c}) + \det(\mathbf{a}|\mathbf{b}|\mathbf{d}).$$

As Figure 2 shows, this too is geometrically desirable. For a two dimensional picture of the same rule see [O, page 9]. Sometimes a three dimensional picture helps bring out the essentially two dimensional nature of the situation!

What is the connection between these ideas? Notice first that if the matrix A has two columns the same, then $\det A = 0$ since the matrix is not invertible or, more geometrically, since the image of the unit cube collapses because two of its edges coincide. Now suppose that we want to swap the first two columns of the matrix $A = (\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3)$. Then because of the additivity

$$\begin{aligned}
0 &= \det(\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3) \\
&= \det(\mathbf{a}_1|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3) \\
&= \det(\mathbf{a}_1|\mathbf{a}_1|\mathbf{a}_3) + \det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_2|\mathbf{a}_3) \\
&= \det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3).
\end{aligned}$$

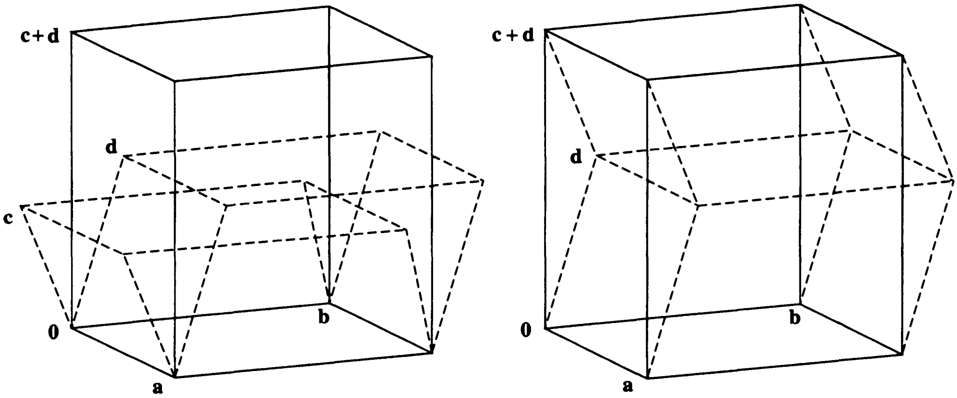


Figure 2. (*Left*) Images of the unit cube under the matrices (**a**|**b**|**c**), (**a**|**b**|**d**) and (**a**|**b**|**c** + **d**). (*Right*) Translating the image corresponding to the matrix (**a**|**b**|**c**) shows that the combined volumes corresponding to the matrices (**a**|**b**|**c**) and (**a**|**b**|**d**) equal the volume corresponding to the matrix (**a**|**b**|**c** + **d**).

Hence

$$\det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) = -\det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3),$$

and so swapping two columns of the matrix changes the sign of the determinant. This means we need to consider the possibility of some volumes being negative when we calculate the magnification factors for det A. This unexpected development forces us to ask: were the pictures misleading?

Not really! It is more a question of us not noticing a subtle feature of the diagrams. Something similar happens when you first meet the integral as a way of calculating areas. There again you begin by imagining all areas are positive. But you have to allow for negative areas when you agree to rules like

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Furthermore, this rule is forced on us (as it was with determinants), if we want an additivity formula

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$$

to hold, even in situations (like $a = c$) that probably were not envisaged when you first tried to calculate areas.

What subtle feature have we missed so far? Clearly a positive determinant must correspond to a particular order of the columns of A. It makes sense that the identity matrix I ought have $\det I = +1$, so we should be able to observe the geometric effect of other column orders by swapping columns in I. Thus, in the 2×2 case, we should compare the effects of I and the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, as in Figure 3.

Although the image still has the same area as the original, it is actually a reflection of the original. We say that the transformation has *reversed the orientation* of the original. In general, a negative determinant indicates that a reflection is part of the action of the matrix.

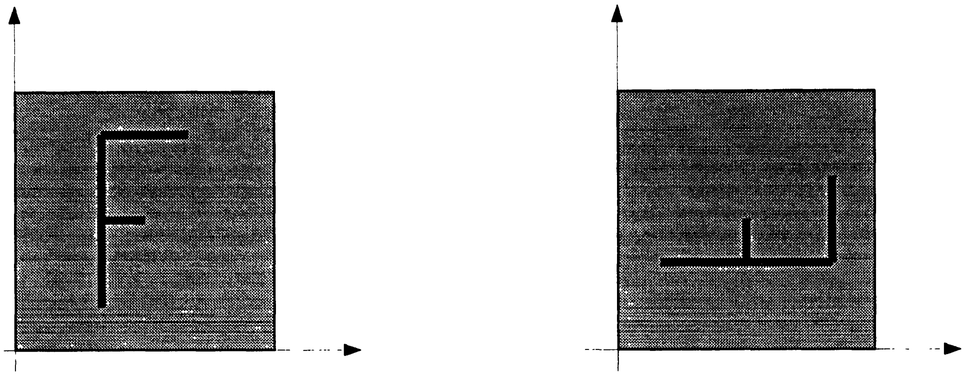


Figure 3. Images of the unit square under the action of the identity matrix (*left*) and the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (*right*).

OTHER COLUMN OPERATIONS

Multiplying a column by a scalar: Since this operation scales one edge of the image, leaving all the other edges constant, the volume is scaled by the same amount. For example,

$$\det(\mathbf{a}_1|s\mathbf{a}_2|\mathbf{a}_3) = s\det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3).$$

Notice that if the scalar is negative, then the orientation of the transformation is reversed.

Adding a multiple of one column to another: This operation produces a shear in the two dimensional picture corresponding to the two affected columns (see Figure 4), so there is no change in the volume or in the orientation. For example,

$$\det(\mathbf{a}|\mathbf{b} + s\mathbf{a}|\mathbf{c}) = \det(\mathbf{a}|\mathbf{b}|\mathbf{c}).$$

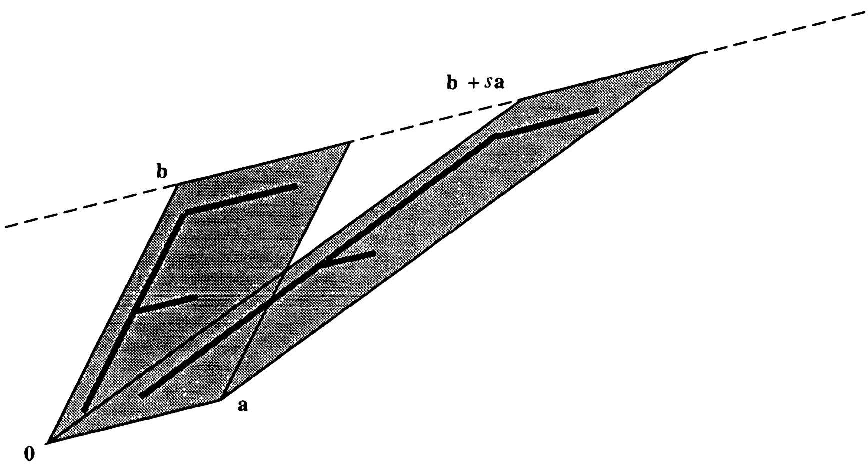


Figure 4. Images of the unit square under the matrices $(\mathbf{a}|\mathbf{b})$ and $(\mathbf{a}|\mathbf{b} + s\mathbf{a})$.

DETERMINANTS AND TRANSPOSES. A remarkable consequence of the product rule is that

$$\det A^T = \det A.$$

At this stage there does not seem to be a simple geometric reason for this result. Except for some simple matrices, there is no obvious geometric relationship between the columns of A and those of A^T. Consequently the explanation offered here will be essentially algebraic. However, as we shall see shortly, once the students have learned about eigenvalues and eigenvectors, we can give genuinely geometric reasons for this formula.

The simplest case is when A is not invertible. Then neither is A^T and both matrices have zero determinant.

So suppose that A is invertible. Then A is a product of elementary matrices. Because of the product rule, it follows that we just need to show that $\det E^T = \det E$ for any elementary matrix E. Now there are three different types of elementary matrix:

1. If E corresponds to the operation of multiplying the ith row by a nonzero scalar s, then E is just the identity matrix with its ith diagonal entry changed to s. So $E = E^T$ and clearly $\det E^T = \det E$.
2. If E corresponds to swapping two rows, then E is got by swapping the same rows of the identity matrix, and again $E = E^T$. So $\det E^T = \det E$ in this case too.
3. Suppose that E corresponds to the operation of adding $s(\text{row } i)$ to row j. This affects only the ith and jth columns of E or E^T, so it is enough to look at what is happening to these two columns in the pictures for the transformations E and E^T. This makes it essentially a 2×2 matrix problem, comparing the determinants of the matrices

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}.$$

In this case, as Figure 5 shows, the corresponding parallelograms are simply reflections of one another. Furthermore, again as in Figure 5, the images still have positive orientation, and so once again $\det E^T = \det E$.

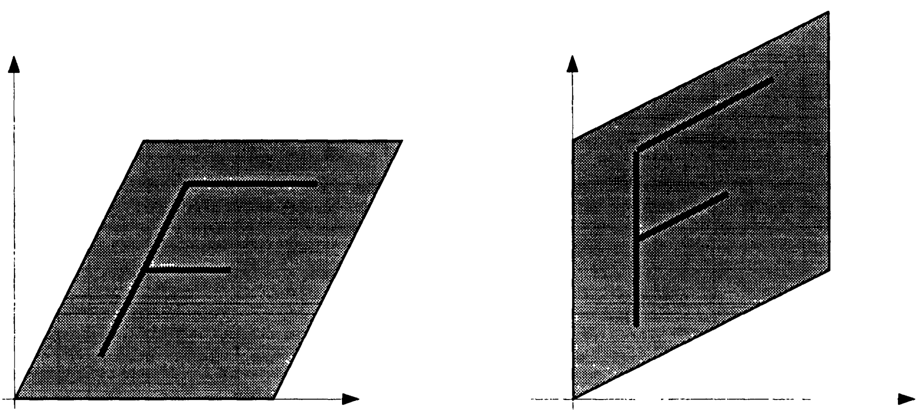


Figure 5. Images of the unit square under the transformations $\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ (*left*) and $\begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}$ (*right*).

CALCULATING THE DETERMINANT. There is nothing new here. We have seen how column operations affect the determinant of a matrix, and we have also seen that $\det A^T = \det A$. Since column operations on A^T are exactly the same thing as row operations on A, this means that we also know the effect of row operations on the determinants. So we can use row operations (or column operations) to reduce the matrix to upper triangular form. Furthermore, we can use the same sort of operations to see that

If A is triangular, then $\det A$ is the product of the entries on the main diagonal of A.

Putting these two steps together gives the most efficient way of calculating $\det A$.

This completes my shopping list of "what every first year student should know about determinants," but of course I have left out several much-loved topics. Peer pressure ensures that I do at least mention them.

FORMULAS. What about a formula for $\det A$? We have now reached the stage where we have enough properties to find out what the formula must be in terms of the entries of A. These calculations also turn up some new ways of finding the determinant: the Laplace expansions of $\det A$ along the various rows and columns of A. See [**St**, pages 223–227] for one way of doing this.

The Laplace expansions of $\det A$ show that the adjugate (or classical adjoint) matrix, $\operatorname{adj} A$, constructed from the cofactors of A, satisfies

$$A \cdot (\operatorname{adj} A) = (\operatorname{adj} A) \cdot A = (\det A) I.$$

Hence as long as $\det A \neq 0$ we have the formula

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A.$$

This formula leads to Cramer's Rule for solving the system of linear equations $A\mathbf{x} = \mathbf{b}$. See [**St**, pages 231–233] for details.

Notice that none of these formulas is of much practical use, except perhaps in the 2×2 case. Their main value lies in their theoretical applications. In practice, $\det A$, A^{-1}, and the solution to $A\mathbf{x} = \mathbf{b}$ should all be found by using row operations (and/or column operations, if that makes sense).

EIGENVALUES, EIGENVECTORS AND DETERMINANTS. Eigenvalues and eigenvectors are another way of looking at the magnification properties of a linear transformation $\mathbf{x} \to A\mathbf{x}$. This time we look at linear magnification (the eigenvalues λ) along those directions (the eigenvectors $\mathbf{x}$) that are preserved by the transformation. Algebraically, these are related by $A\mathbf{x} = \lambda\mathbf{x}$.

In the case where the matrix A is diagonalizable (as is the case when the real matrix A is symmetric, for example), this gives another picture of the effect of A on volumes. Figure 6 shows a typical situation in the case where A is a 2×2 matrix.

So, at least in the diagonalizable case, we see that

$$\det A = \text{the product of the eigenvalues of } A.$$

Anecdote: Last time I taught this topic, I drew the picture in Figure 6 to show how the eigenvalues and eigenvectors could be used to describe the whole transformation. I did not mention determinants as I did not need the preceding displayed formula for later work in the course. However, one student came to me at the end of the session and asked whether the picture meant that $\det A$ had to be the product of the eigenvalues of A. The algebraic proof of this fact is simple,

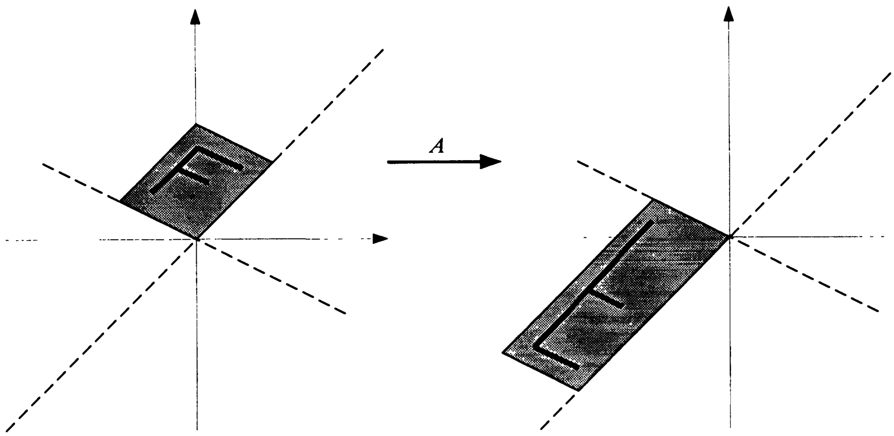


Figure 6. Effect of a matrix A with eigenvalues $\lambda = 1, -2$ and eigenvectors parallel to the dashed lines.

too (in the diagonalizable case), yet I've never had a student make the same connection when I have given an algebraic treatment of determinants.

DETERMINANTS AND VOLUMES OF ELLIPSOIDS. An n-dimensional ellipsoid is determined by an equation of the form $\mathbf{x}^T M \mathbf{x} = 1$ where M is a positive definite (real symmetric) matrix. The determinant can be used to find the (positive!) volume of this ellipsoid. Using a suitable rotation of axes we may assume that M is a diagonal matrix without altering $\det M$. So the equation of the ellipsoid is

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_n x_n^2 = 1$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of M. This is the same as

$$\frac{x_1^2}{\left(1/\sqrt{\lambda_1}\right)^2} + \frac{x_2^2}{\left(1/\sqrt{\lambda_2}\right)^2} + \cdots + \frac{x_n^2}{\left(1/\sqrt{\lambda_n}\right)^2} = 1,$$

which corresponds to a unit n-sphere that has been scaled by $1/\sqrt{\lambda_1}$ in the first coordinate direction, by $1/\sqrt{\lambda_2}$ in the second coordinate direction, and so on. Hence the volume of the ellipsoid must be

$$\left(\frac{1}{\sqrt{\lambda_1}}\right) \cdots \left(\frac{1}{\sqrt{\lambda_n}}\right) V = \frac{V}{\sqrt{\lambda_1 \cdots \lambda_n}} = \frac{V}{\sqrt{\det M}},$$

where V is the volume of the unit n-ball. For example, in the 2×2 case, where the unit disk has area π, the ellipse determined by $\mathbf{x}^T M \mathbf{x} = 1$ has area $\pi/\sqrt{\det M}$. Similarly, in the 3×3 case, the ellipsoid determined by $\mathbf{x}^T M \mathbf{x} = 1$ has volume $\frac{4}{3}\pi/\sqrt{\det M}$.

A GEOMETRIC VIEW OF THE DETERMINANT OF THE TRANSPOSE. We are now ready to see a more geometric explanation of the rule

$$\det A^T = \det A.$$

We can assume that A is an invertible matrix. The key idea is to evaluate the volume magnification of the transformation $\mathbf{x} \to A\mathbf{x}$ by looking at its effect on the unit n-sphere. It ends up being simpler to find the *inverse* image of the unit n-sphere $\mathbf{z}^T \mathbf{z} = 1$ (Figure 7).

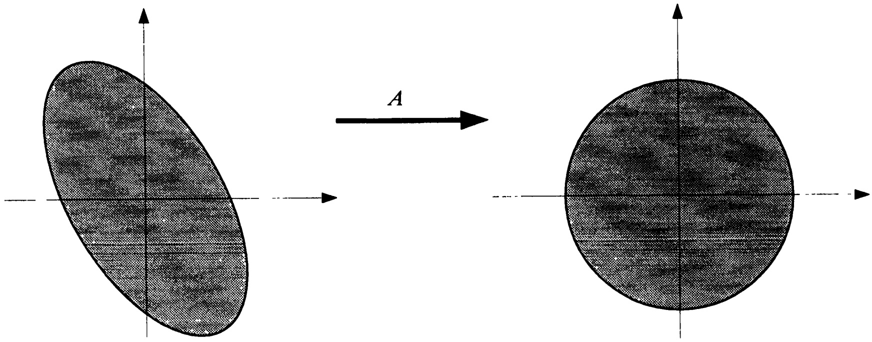


Figure 7. The inverse image under A of the unit n-sphere.

The points $\mathbf{x}$ mapping onto this sphere clearly come from the ellipsoid

$$1 = (A\mathbf{x})^T A\mathbf{x} = \mathbf{x}^T(A^TA)\mathbf{x}.$$

Now A^TA is a positive definite matrix, so the preceding discussion on volumes of ellipsoids can be applied here. Thus the inverse image has (positive) volume

$$\frac{V}{\sqrt{\det A^TA}}$$

where V is again the volume of the unit n-ball. Hence the (signed) magnification factor of the transformation $\mathbf{x} \to A\mathbf{x}$ is given by $\pm\sqrt{\det A^TA}$. By our definition of $\det A$ we thus have $\det A = \pm\sqrt{\det A^TA}$. Squaring and using the product rule to cancel a factor of $\det A$, we get the desired formula.

REFERENCES

[AK] Agnew, J. L., and Knapp, R. C., *Linear algebra with applications*, 3rd ed., Brooks/Cole, 1989.

[BK] Blum, W., and Kirsch, A., Preformal proving: examples and reflections, *Educ. Stud. Math.* **22** (1991) 183–203.

[DL] Damiano, D. B., and Little, J. B., *A course in linear algebra*, Harcourt, Brace and Jovanovich, 1988.

[FB] Fraleigh, J. B., and Beauregard, R. A., *Linear Algebra*, 2nd ed., Addison-Wesley, 1990.

[N] Nicholson, W. K., *Elementary linear algebra with applications*, 2nd ed., PWS-Kent, 1990.

[O] Open University, *Determinants and eigenvalues*, (Linear Mathematics, Unit 5), The Open University Press, 1972.

[Se] Sernesi, E., *Linear algebra: a geometric approach*, Chapman and Hall, 1993.

[St] Strang, G., *Linear Algebra and its Applications* (3rd edition), Harcourt Brace Jovanovich, 1988.

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch
New Zealand
j.hannah@math.canterbury.ac.nz

NOTES

Edited by: **John Duncan**

On Zero Derivatives

Darrell Desbrow

In complex analysis, the fact that zero derivatives imply constancy of values, either locally or globally on domains, is often proved by invoking the Cauchy-Riemann equations to show that the partial derivatives of both the real and imaginary parts are similarly zero. By that or other means, the matter is reduced to a corresponding result of real calculus, *viz.*, that zero derivatives imply constancy of values on intervals. The traditional proof of that result relies upon the Mean Value Theorem, although several proofs are known that avoid using it; see Richmond [3] for a recent example of such a proof, as well as references to earlier real-calculus proofs. There is an earlier proof along similar lines by Cohen [2]. The reader is particularly referred to Bers' proof in the real-valued case [1], which is 'from first principles' and hence shares at least that much with the proof we offer.

To a novice at least, all this looks like something of a *tour de force* for what seems to be a result ripe for proof directly from definition of the derivative. Such a proof is given here, and an elementary one, which applies equally well whether the function values or variable are real or complex (or rather more general). The result is stated here in the traditional complex analysis version only but it can be easily particularized to real values or variables (or indeed generalized to Banach spaces), as desired.

Theorem. *Let f be a complex-valued function whose derivative is everywhere zero on the disc $D: |z - a| < r$, for some complex number a and positive real number r. Then f is constant on D.*

It follows that a function whose derivative is zero everywhere on, respectively, an open set/a domain is locally constant/constant.

By considering instead the function g defined by $g(z) = f(z + a) - f(a)$, which has zero derivative at z when f has the same at $z + a$, it may be assumed that $a = 0 = f(a)$. It must then be shown that f vanishes with its derivative everywhere on D: $|z| < r$. This we do contrapositively.

Proof: Suppose that $f(0) = 0$ and f is differentiable but not everywhere zero on the disc D: $|z| < r$. Then $|f(w)| = C|w|$ for some non-zero $w \in D$ and $C > 0$. Let

$$A = \{t: 0 < t \leq 1, |f(tw)| \geq C|tw|\}.$$

Then $1 \in A$ and so A has a greatest lower bound, λ say, with $0 \leq \lambda \leq 1$. Clearly $|f(\lambda w)| \geq C|\lambda w|$.

Case $\lambda = 0$: There is a sequence (t_n) in A with $t_n \to \lambda = 0$. Thus for each n,

$$\left|\frac{f(t_n w)}{t_n w}\right| \geq C.$$

Letting $n \to \infty$, so that $t_n w \to 0$, we conclude that $|f'(0)| \geq C > 0$. Thus $f'(0) \neq 0$.

Case $\lambda > 0$: For any μ with $0 < \mu < \lambda$, if $|f(\mu w) - f(\lambda w)| \leq C|\mu w - \lambda w|$ then

$$\begin{aligned}|f(\mu w)| &\geq |f(\lambda w)| - |f(\mu w) - f(\lambda w)| \\ &\geq C\lambda|w| - C(\lambda - \mu)|w| = C|\mu w|,\end{aligned}$$

contradicting the definition of λ. Hence when $0 < \mu < \lambda$,

$$\left|\frac{f(\mu w) - f(\lambda w)}{\mu w - \lambda w}\right| > C.$$

Letting $\mu \to \lambda$ from below, so that $\mu w \to \lambda w$, we conclude that $|f'(\lambda w)| \geq C > 0$. Thus $f'(\lambda w) \neq 0$.

In either case, the result is proved.

REFERENCES

1. Bers, L., On avoiding the mean value theorem, *Amer. Math. Monthly* **74** (1967), 583.
2. Cohen, L. W., On being mean to the mean value theorem, *Amer. Math. Monthly* **74** (1967), 581–582.
3. Richmond, D. E., An elementary proof of a theorem of calculus, *Amer. Math. Monthly* **92** (1985), 589–590.

Department of Mathematics & Statistics
University of Edinburgh
Edinburgh
SCOTLAND
EH9 3JZ.
E-mail: dd@maths.ed.ac.uk

The Argument Principle for Harmonic Functions

Peter Duren, Walter Hengartner, and Richard S. Laugesen

The argument principle is one of the great "miracles" of analytic function theory, and has many important applications. Amid a recent surge of interest [**1**] in planar harmonic mappings, or complex-valued univalent harmonic functions (not necessarily analytic), there has been good reason to look for a direct and simple "harmonic generalization" of the principle. More broadly generalized versions have indeed been available for some time but have made use of heavy tools such as quasiconformal mappings.

The purpose of this note is to develop a precise and *elementary* generalization of the argument principle to harmonic functions, at the level of a first course in complex analysis.

Let us first recall the argument principle for analytic functions and its elegant proof. Let D be a plane domain bounded by a rectifiable Jordan curve C, oriented in the positive or "counterclockwise" direction. Let f be analytic in D and continuous in $\overline{D}$, with $f(z) \neq 0$ on C. The *index* or *winding number* of the image curve $f(C)$ about the origin is then defined to be $I = (1/2\pi)\Delta_C \arg f(z)$, the net change in the argument of $f(z)$ as z runs once around C, divided by 2π. Let N be the total number of zeros of f in D, counted according to multiplicity. Then the argument principle asserts that $N = I$.

The customary proof begins with the observation that f'/f has a simple pole with residue n wherever f has a zero of order n, so the residue theorem gives

$$N = \frac{1}{2\pi i}\int_C \frac{f'(z)}{f(z)}\,dz = \frac{1}{2\pi i}\,\Delta_C \log f(z) = I.$$

(Actually, since the derivative $f'(z)$ need not be defined on C, the contour of integration should be slightly contracted.)

The argument principle is a practical tool for deducing analytic information from more apparent geometric behavior. As a typical application, it can be seen that if f is analytic in D and continuous in $\overline{D}$, and if it carries C univalently in a sense-preserving manner onto a Jordan curve Γ bounding a domain Ω, then f maps D univalently onto Ω. In other words, univalence on the boundary implies univalence in the interior.

In order to formulate the argument principle for harmonic functions, one must first make sense of the notion of "order of a zero". A *harmonic function* is a solution of Laplace's equation $f_{xx} + f_{yy} = 0$. Every harmonic function $f = u + iv$ has a local decomposition $f = h + \bar{g}$ in terms of two analytic functions h and g, which are unique up to additive constants. (To see this, observe first that $f_z = \frac{1}{2}(f_x - if_y)$ satisfies the Cauchy-Riemann equations and is therefore analytic. Now define $h' = f_z$, integrate to obtain h, and show that $g = \bar{f} - \bar{h}$ also satisfies the Cauchy-Riemann equations.) An elementary calculation gives the Jacobian of f as $J_f = u_x v_y - u_y v_x = |h'|^2 - |g'|^2$. A harmonic function f will be called *sense-preserving* at a point z_0 if $h'(z) \not\equiv 0$ and $\omega = g'/h'$ is analytic at z_0 (possibly with a removable singularity), and $|\omega(z_0)| < 1$. In justification of this definition, notice that if f is sense-preserving then $J_f(z_0) > 0$ unless $h'(z_0) = 0$. Similarly, f is *sense-reversing* at z_0 if $\bar{f} = g + \bar{h}$ is sense-preserving there. Call z_0 a *singular point* if f is neither sense-preserving nor sense-reversing at z_0. It is clear that $J_f(z_0) = 0$ at every singular point, but not conversely. Examples of sense-preserving harmonic functions are all nonconstant analytic functions, and the functions $\alpha z^n + \beta \bar{z}^m$ for $|z| < 1$, with $n \le m$ and $m|\beta| < n|\alpha|$. Every point z_0 is singular for the function $z + \bar{z}$.

For a sense-preserving harmonic function f, the order of a zero can be defined in terms of the local decomposition $f = h + \bar{g}$. Suppose $f(z_0) = 0$ at some point z_0 where f is sense-preserving, and write the power-series expansions of h and g as

$$h(z) = a_0 + \sum_{k=1}^{\infty} a_k (z - z_0)^k, \qquad g(z) = b_0 + \sum_{k=1}^{\infty} b_k (z - z_0)^k.$$

Actually, $b_0 = -\overline{a_0}$ because $f(z_0) = 0$. Some a_k $(k \ge 1)$ must be nonzero because $h'(z) \not\equiv 0$. Let a_n be the first such nonzero coefficient. Then $b_k = 0$ for $1 \le k < n$,

since $\omega = g'/h'$ is analytic at z_0; and $|b_n| < |a_n|$ because $|\omega(z_0)| < 1$. We will say that f has a zero of *order n* at z_0. Similarly, if f is sense-reversing at a zero z_0, then $\bar{f}$ is sense-preserving at z_0 with a zero of some order n, and we say that f has a zero of *order* $-n$ at z_0. The order of a singular zero is not defined.

As an immediate consequence of the structural formulas just derived, it can be inferred that the nonsingular zeros of a harmonic function are isolated. Indeed, if $f(z_0) = 0$ and $|\omega(z_0)| < 1$, then for $0 < |z - z_0| < \delta$ it is possible to write

$$f(z) = h(z) + \overline{g(z)} = a_n(z - z_0)^n\{1 + \psi(z)\},$$

where

$$\psi(z) = (\overline{b_n}/a_n)(\bar{z} - \bar{z}_0)^n(z - z_0)^{-n} + O(z - z_0).$$

But it is clear that $|\psi(z)| < 1$ for z sufficiently close to z_0, since $|b_n/a_n| < 1$. Hence $f(z) \neq 0$ near z_0. A similar argument applies at the sense-reversing zeros of f. Thus each nonsingular zero of f is isolated. However, the singular zeros of a harmonic function are not always isolated. For example, $f(z) = z + \bar{z} = 2x$ vanishes on the whole imaginary axis.

The argument principle for harmonic functions can now be formulated as a direct generalization of the classical result for analytic functions.

Theorem. *Let f be a harmonic function in a Jordan domain D with boundary C. Suppose f is continuous in $\overline{D}$ and $f(z) \neq 0$ on C. Suppose f has no singular zeros in D, and let N be the sum of the orders of the zeros of f in D. Then $\Delta_C \arg f(z) = 2\pi N$.*

Proof: Suppose first that f has no zeros in D, so that $N = 0$ and the origin lies outside $f(D \cup C)$. A fact from topology says that in this case $\Delta_C \arg f(z) = 0$, which proves the theorem. To prove the topological fact, let ϕ be a homeomorphism of the closed unit square S onto $D \cup C$ with ϕ: $\partial S \to C$ a homeomorphism. Then $F = f \circ \phi$ is a continuous map of S into the plane with no zeros, and we want to prove that $\Delta_{\partial S} \arg F(z) = 0$. Begin by subdividing S into finitely many small squares S_j on each of which the argument of $F(z)$ varies by at most $\pi/2$. Then $\Delta_{\partial S_j} \arg F(z) = 0$ and so

$$\Delta_{\partial S} \arg F(z) = \sum_j \Delta_{\partial S_j} \arg F(z) = 0,$$

the first equality relying on the cancellation of contributions from the ∂S_j except on ∂S.

Next suppose that f does have zeros in D. Because the zeros are isolated and f does not vanish on C, there are only a finite number of distinct zeros in D. Denote them by z_j for $j = 1, 2, \dots, \nu$. Let γ_j be a circle of radius $\delta > 0$ centered at z_j, where δ is chosen so small that the circles γ_j all lie in D and do not meet each other. Join each circle γ_j to C by a Jordan arc λ_j in D. Consider the closed path Γ formed by moving around C in the positive direction while making a detour along each λ_j to γ_j, running once around this circle in the negative (clockwise) direction, then returning along λ_j to C. The curve Γ contains no zeros of f, and so $\Delta_\Gamma \arg f(z) = 0$ by the case considered above. But the contributions of the arcs λ_j along Γ cancel out, so that

$$\Delta_C \arg f(z) = \sum_{j=1}^{\nu} \Delta_{\gamma_j} \arg f(z),$$

where now each of the circles γ_j is traversed in the positive direction. This formula reduces the global problem to a local one. (The same reduction is often used to prove the residue theorem.)

Suppose now that f has a zero of order $n > 0$ at a point z_0. Then as observed above, f has the local form $f(z) = a_n(z - z_0)^n\{1 + \psi(z)\}$, $a_n \neq 0$, where $|\psi(z)| < 1$ on a sufficiently small circle γ defined by $|z - z_0| = \delta$. This shows that

$$\Delta_\gamma \arg f(z) = n\,\Delta_\gamma \arg\{z - z_0\} + \Delta_\gamma \arg\{1 + \psi(z)\} = 2\pi n.$$

Similarly, the same conclusion $\Delta_\gamma \arg f(z) = 2\pi n$ holds if f has a zero of order $n < 0$ at z_0. Therefore, if f has zeros of order n_j at the points z_j, the conclusion is that

$$\Delta_C \arg f(z) = \sum_{j=1}^{\nu} \Delta_{\gamma_j} \arg f(z) = 2\pi \sum_{j=1}^{\nu} n_j = 2\pi N,$$

which proves the theorem. There is an obvious extension to multiply connected domains, as for analytic functions.

Several corollaries are worthy of note. First of all, there is a direct extension of Rouché's theorem to harmonic functions. Specifically, if p and $p + q$ are sense-preserving harmonic functions in D, continuous in $\overline{D}$, and if $|q(z)| < |p(z)|$ on C, then p and $p + q$ have the same number of zeros inside D. As in the proof for analytic functions, the strict inequality on C implies that neither p nor $p + q$ has a zero on C and that the images of C under the two functions have the same winding number about the origin. Thus the harmonic version of Rouché's theorem follows from the harmonic version of the argument principle. In fact, the theorem remains true and the proof is the same if the "sense-preserving" hypothesis on p and $p + q$ is relaxed to the requirement that they have no singular zeros, but then the "number of zeros" must be interpreted as the sum of the (positive or negative) orders of the zeros.

Next there is a generalization of Hurwitz's theorem. If f_k are harmonic functions (in a domain D) that converge locally uniformly, then their limit function f is harmonic. The "harmonic" Hurwitz theorem claims that if f and all the f_k are sense-preserving, then a point z_0 in D is a zero of f if and only if it is a cluster-point of zeros of the functions f_k. More precisely, f has a zero of order n at z_0 if and only if each small disk around z_0 (small enough to contain no other zeros of f) contains precisely n zeros, counted according to multiplicity, of f_k for every k sufficiently large. The proof applies Rouché's theorem exactly as in the analytic case, with $p = f$ and $q = f_k - f$.

Finally, sense-preserving harmonic functions have the *open mapping property*: they carry open sets to open sets. In fact, as in the analytic case, a stronger statement can be made. If f is a sense-preserving harmonic function near a point where $f(z_0) = w_0$, and if $f(z) - w_0$ has a zero of order n ($n \geq 1$) at z_0, then to each sufficiently small $\varepsilon > 0$ there corresponds a $\delta > 0$ with the following property. For each $\alpha \in N_\delta(w_0) = \{w\colon |w - w_0| < \delta\}$, the function $f(z) - \alpha$ has exactly n zeros, counted according to multiplicity, in $N_\varepsilon(z_0)$. The proof appeals to Rouché's theorem with $p = f - w_0$ and $q = w_0 - \alpha$.

ACKNOWLEDGMENTS. Duren's and Laugesen's work were supported in part by NSF Grants DMS-9401693 and DMS-9414149; Hengartner's by grants from NSERC and FCAR.

REFERENCE

1. Duren, P. "A survey of harmonic mappings in the plane", Texas Tech University, Mathematics Series, Visiting Scholars' Lectures 1990–1992, vol. 18 (1992), pp. 1–15.

Peter Duren and Richard S. Laugesen
Department of Mathematics
University of Michigan
Ann Arbor, Michigan 48109
U.S.A.
e-mail: duren@math.lsa.umich.edu,
laugesen@math.lsa.umich.edu

Walter Hengartner
Département de Mathématiques
Université Laval
Québec, P.Q. G1K 7P4
CANADA
e-mail: walheng@mat.ulaval.ca,

Differentiating Powers of an Old Friend

Richard B. Darst and Gerald D. Taylor

The function $r: R \to R$ defined by $r(x) = x$ when x is rational and $r(x) = 0$ otherwise is continuous only at $x = 0$. Changing the non-zero rational values from m/n to $1/n$ when $x = m/n$ with m and n relatively prime integers and $n > 0$ defines an old friend, f, which is continuous at zero and at all irrational numbers. Powers, f^t, of f are defined for $t > 1$ by $f^t(x) = [f(x)]^t$. Notice that f is nowhere differentiable; differentiability properties of f^t follow.

Theorem. *If* $1 < t \leq 2$, *then* f^t *is differentiable only at zero. If* $t > 2$, *then* f^t *is differentiable almost everywhere.*

Proof: The case where $t \leq 2$ follows from Theorem 4.1 in [**1**]; consequently, we consider the case $t = 2 + 2p$ where $p > 0$ and observe that when x is an irrational,

$$\left| \frac{f^t(x) - f^t(m/n)}{x - m/n} \right| = \frac{1/n^{2+2p}}{|x - m/n|} \leq 1/n^p$$

if

$$|x - m/n| \geq 1/n^{2+p}.$$

So we put

$$E_n = \bigcup_{1 \leq m < n} (m/n - 1/n^{2+p}, m/n + 1/n^{2+p}),$$

and let

$$E = \bigcap_{k \geq 1} \bigcup_{n \geq k} E_n.$$

Since $1/n^p \to 0$ as $n \to \infty$, the derivative of f' is zero at an irrational number that is in $(0,1)$ but not in E. Because the Lebesgue measure of E_n is equal to $2(n-1)/n^{2+p} < 2/n^{1+p}$ and $\Sigma_{n \geq 1} 1/n^{1+p} < \infty$, the Lebesque measure of E is zero.

Translation mod 1 completes the proof.

REFERENCE

1. Niven, Ivan, Irrational Numbers. The Carus Mathematical Monographs, Number 11, Mathematical Association of America, Washington, D.C., 1967.

Richard B. Darst and Gerald D. Taylor
Department of Mathematics
Colorado State University
Fort Collins, CO 80523-0001
email: darst@math.colostate.edu

God is a child; and when He began to play, He cultivated mathematics. It is the most godly of man's games.

Vinzenz Erath
Das Blinde Spiel, Wunderlich, Tübingen, 1954

Mathematics is like checkers in being suitable for the young: not too difficult, amusing, and without peril to the state.

Plato (?)

THE EVOLUTION OF ...

Edited by: **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Genesis of the Abstract Ring Concept

Israel Kleiner

Algebra textbooks usually give the definition of a ring first and follow it with examples. Of course, the examples came first, and the abstract definition later—much later. So we begin with examples.

Among the most elementary examples of rings are the integers, polynomials, and matrices. "Simple" extensions of these examples are at the roots of ring theory. Specifically, we have in mind the following three examples:

(a) The integers Z can be thought of as the appropriate subdomain of the field Q of rationals in which to do number theory. (The rationals themselves are unsuitable for that purpose: every rational is divisible by every other (nonzero) rational.) Take a simple extension field $Q(\alpha)$ of the rationals, where α is an algebraic number, that is, a root of a polynomial with integer coefficients. $Q(\alpha)$ is called an algebraic number field; it consists of polynomials in α with rational coefficients (e.g., $Q(\sqrt{3}) = \{a + b\sqrt{3}: a, b \in Q\}$). The appropriate subdomain of $Q(\alpha)$ in which to do number theory—the "integers" of $Q(\alpha)$—consists of those elements that are roots of *monic* polynomials with integer coefficients (the integers of $Q(\sqrt{3})$ are $\{a + b\sqrt{3}: a, b \in Z\}$). This is our first example.

(b) The polynomial rings $\mathbb{R}[x]$ and $\mathbb{R}[x, y]$ in one and in two variables, respectively, share important properties but also differ in significant ways. In particular, while the roots of a polynomial in one variable constitute a discrete set of real numbers, the roots of a polynomial in two variables constitute a curve in the plane—a so-called algebraic curve. Our second example, then, is the ring of polynomials in two (or more) variables.

(c) Square $m \times m$ matrices (for example, over the reals) can be viewed as m^2-tuples of real numbers with coordinate-wise addition and appropriate multiplication obeying the axioms of a ring. Our third example consists, more generally, of n-tuples $\mathbb{R}^n$ of real numbers with coordinate-wise addition and appropriate multiplication, so that the resulting system is a (not necessarily commutative) ring. Such systems are extensions of the complex numbers—in the 19th and early 20th centuries they were called hypercomplex number systems.

In what contexts did these examples arise? What was their importance? The answers will lead us to the genesis of the abstract ring concept.

The abstract ring concept emerged in the context of a theory—in fact, in the context of two theories: commutative ring theory and noncommutative ring theory.

The abstract theories of these two categories came from distinct sources and developed in different directions. Commutative ring theory originated in algebraic number theory and algebraic geometry. Central to the development of these subjects were, respectively, the rings of integers in algebraic number fields and the rings of polynomials in two or more variables. Noncommutative ring theory began with attempts to extend the complex numbers to various hypercomplex number systems. We consider first the evolution of the "simpler" theory of noncommutative rings.

A. NONCOMMUTATIVE RING THEORY. In a strict sense, noncommutative ring theory originated from a single example—the quaternions—invented (discovered?) by Hamilton in 1843. These are "numbers" of the form $a + bi + cj + dk$ (a, b, c, d real numbers) that are added componentwise and in which multiplication is subject to the relations $i^2 = j^2 = k^2 = ijk = -1$. This was the first example of a noncommutative number system, obeying all the (algebraic) laws of the real and complex numbers except for commutativity of multiplication. Such a system is now called a skew field or a division algebra. Hamilton's motivation was to define an algebra of vectors in 3-space so that multiplication would represent composition of rotations (just as multiplication of complex numbers represents composition of rotations in the plane). Having failed in this task, he turned to quadruples of reals and created the algebra of quaternions. The "pure" quaternions did, in fact, yield the required computing tool for rotations in 3-space.

Examples. Hamilton's invention of the quaternions was conceptually groundbreaking, but like all revolutions, it was initially received with less than universal approbation. Most mathematicians, however, soon came around to Hamilton's point of view. The quaternions acted as a catalyst for the exploration of diverse "number systems", with properties that departed in various ways from those of the real and complex numbers. Among the examples of such hypercomplex number systems are octonions, exterior algebras, group algebras, matrices, and biquaternions. See **[6]**.

Structure. The first example of a noncommutative algebra was given by Hamilton in 1843. During the next forty years mathematicians introduced other examples, began to bring some order into them, and singled out certain types for special attention. For example, Frobenius and C. S. Peirce showed in 1880 that the reals, the complex numbers, and the quaternions are the only finite-dimensional (associative) division algebras over the reals. The stage was set for the founding of a general theory of finite-dimensional, noncommutative, associative algebras (important examples of rings).

In the 1890s, Cartan, Frobenius, and Molien proved (independently) the following fundamental structure theorem for finite-dimensional semi-simple algebras over the real or complex numbers: Any such algebra is a finite unique direct sum of simple algebras. These, in turn, are isomorphic to matrix algebras with entries in division algebras. An algebra is "semi-simple" if it has no nontrivial nilpotent ideals and it is "simple" if it has no nontrivial ideals.

In 1907, Wedderburn extended that result to algebras over arbitrary fields. This was no mere generalization as it necessitated the introduction of such fundamental algebraic concepts as ideal, quotient algebra, nilpotent algebra, radical, semi-simple and simple algebra, direct sum, and tensor product. Wedderburn's theorem is one of the basic results in ring theory, and it serves as a model for many

ring-theoretic structure theorems. It is also central in group-representation theory. See [**6**].

B. COMMUTATIVE RING THEORY. *Commutative* ring theory originated in algebraic number theory and algebraic geometry and has in turn been applied mainly to these two subjects. Invariant theory, with roots in both number theory and geometry, also had a role in these developments.

Algebraic number theory. Several of the central areas of number theory, principally Fermat's Last Theorem, binary quadratic forms, and reciprocity laws, were instrumental in the emergence of algebraic number theory. Although the key problems in these areas were expressed in terms of integers, it gradually became apparent that the solutions called for embedding the integers in domains of what came to be known as algebraic integers. The following examples give an idea of what is involved.

(i) To show that $x^3 + y^3 = z^3$ has no nonzero integer solutions, factor the left side as $(x + y)(x + yw)(x + yw^2) = z^3$, $w^3 = 1$, $w \neq 1$. This is now an equation in the domain $D_3 = \{a + bw\colon a, b \in Z\}$. Assuming the existence of a solution in D_3 one can arrive at a contradiction, showing in particular that $x^3 + y^3 = z^3$ has no solutions. A similar approach turned out to be fruitful in the general case of Fermat's Last Theorem, $x^p + y^p = z^p$. See [**5**].

(ii) A conceptual way to determine which integers are sums of two squares is to factor the right side of $n = x^2 + y^2$ and consider the equation $n = (x + yi)(x - yi)$ in the domain $G = \{a + bi\colon a, b \in Z\}$ of Gaussian integers. (Gauss introduced G in order to *state* the biquadratic reciprocity law.) This problem is but an instance of the problem of representing integers by binary quadratic forms $ax^2 + bxy + cy^2$ $(a, b, c \in Z)$. The general approach is to factor $ax^2 + bxy + cy^2$ and to consider the resulting equation in a domain of "complex integers". See [**1**].

(iii) The diophantine equation $x^2 + 2 = y^3$ is a special case of the famous Bachet equation $x^2 + k = y^3$. It was in the margins of Bachet's Latin translation of Diophantus' *Arithmetica* that Fermat made his famous remark about the equation $x^n + y^n = z^n$. While the general Bachet equation is still a topic of intensive investigation, the equation $x^2 + 2 = y^3$ can readily be solved using "complex integers". It is easy to see that $x = \pm 5$, $y = 3$ are solutions. To find *all* solutions, we write $x^2 + 2 = y^3$ as $(x + \sqrt{2}i)(x - \sqrt{2}i) = y^3$. This is now an equation in the domain $D = \{a + b\sqrt{2}i\colon a, b \in Z\}$. We can show that D is a unique factorization domain (UFD) and that $x + \sqrt{2}i$ and $x - \sqrt{2}i$ are relatively prime in D. Since their product is a cube, each factor must be a cube (in D). In particular, $x + \sqrt{2}i = (a + b\sqrt{2}i)^3$, $a, b \in Z$. Cubing and then equating coefficients, we can easily show that $x = \pm 5$, $y = 3$ are the *only* solutions of $x^2 + 2 = y^3$—no easy feat to accomplish without the use of complex integers. See [**1**].

What is common to these examples? Additive problems in Z have been transformed to multiplicative problems in the domains D_3, G, D respectively (these domains are important examples of rings). In the latter settings the problems can be dealt with effectively provided that the domains in question are unique

factorization domains (UFDs). Now, D_3, G, and D *are* UFDs, but the domains arising from the respective general problems (Fermat's Last Theorem, binary quadratic forms, the Bachet equation) are (as a rule) not. For example, $\{a_0 + a_1 w + a_2 w^2 + \cdots + a_{22} w^{22}$: $a_i \in Z$, w a primitive 23rd root of unity$\}$, arising from the equation $x^{23} + y^{23} = z^{23}$, is not a UFD; neither is the domain $\{a + b\sqrt{5}i\colon a, b \in Z\}$ resulting from factoring the left side of the Bachet equation $x^2 + 5 = y^3$. The problem then becomes one of restoring, *in some sense*, unique factorization in such domains. Kummer dealt with it by means of ideal numbers, Dedekind by means of ideals, and Kronecker by means of divisors. We consider briefly Dedekind's contribution.

Dedekind's rings and ideals. The main result of Dedekind's groundbreaking 1871 work, which appeared as Supplement X of Dirichlet's *Vorlesungen über Zahlentheorie*, was that every nonzero ideal in the domain of integers of an algebraic number field is a unique product of prime ideals. Before one could state this theorem one had, of course, to define the concepts in its statement, namely "the domain of integers of an algebraic number field", "ideal", and "prime ideal". It took Dedekind about twenty years to formulate them.

Given an algebraic number field $Q(\alpha)$, all its elements are roots of polynomials with integer coefficients. Dedekind defined the *domain of integers* of $Q(\alpha)$ to be the subset of elements that are roots of *monic* polynomials with integer coefficients. This notion is an extension of the domain of integers (of Q), whose elements are the roots of monic *linear* polynomials. He showed that these elements "behave" like integers—they are closed under addition, subtraction, and multiplication. For example, the integers of $Q(\sqrt{5}) = \{a + b\sqrt{5}\colon a, b \in Q\}$ are $\{\frac{a + b\sqrt{5}}{2}\colon a, b \in Z,\ a \equiv b (\operatorname{mod} 2)\}$ rather than $\{a + b\sqrt{5} = a, b \in Z\}$, as seems perhaps more natural. See [1].

Having defined the domain R of algebraic integers of $Q(\alpha)$ in which he would formulate and prove his result on unique decomposition of ideals, Dedekind considered, more generally, sets of integers of $Q(\alpha)$ closed under addition, subtraction, and multiplication. He called these "orders". The domain R of integers of $Q(\alpha)$ is the largest order. Here, then, was an algebraic first for Dedekind—an essentially axiomatic definition of a (commutative) ring, albeit in a concrete setting. The *term* ring, also in the setting of domains of algebraic integers, was coined by Hilbert in 1897.

The second fundamental concept of Dedekind's theory, that of ideal, derived its motivation (and name) from Kummer's ideal numbers (see [2]). Dedekind defined it essentially as we do today. Having then defined the notion of prime ideal, he proved his fundamental theorem that every nonzero ideal in the ring of integers of an algebraic number field is a unique product of prime ideals. See [2].

How did Dedekind's theory relate to the number-theoretic problems (e.g., Fermat's Last Theorem, reciprocity laws, binary quadratic forms) from which it drew inspiration? It did shed important light on these problems and resolved special cases (see [1], [5]). But as often happens, the ideas Dedekind put forth acquired great significance independent of the original problems that stimulated their introduction. (Galois theory also far superseded in importance the problem of solution of equations that gave it birth.)

Algebraic geometry. Algebraic geometry is the study of algebraic curves and their generalizations to n dimensions, algebraic varieties. An algebraic curve is the set of

roots of an algebraic function; that is, a function $y = f(x)$ defined implicitly by the polynomial equation $P(x, y) = 0$.

Several approaches were used in the study of algebraic curves, notably the analytic, the geometric-algebraic, and the algebraic-arithmetic. In the analytic approach, to which Riemann (in the 1850s) was the major contributor, the main objects of study were algebraic functions $f(w, z) = 0$ (of a complex variable) and their integrals, the so-called abelian integrals, which are closely related to the important notion of the genus of an algebraic curve. It was in this connection that Riemann introduced the fundamental notion of a Riemann surface, on which algebraic functions become single valued. Riemann's methods were brilliant but nonrigorous, and relied heavily on the physically obvious Dirichlet Principle, which was mathematically incorrect in its unrestricted form.

(i) Algebraic function fields. Dedekind and Weber, in their groundbreaking 1882 paper "Theory of algebraic functions of a single variable", proposed to "provide a basis for the theory of algebraic functions, the major achievement of Riemann's researches, in the simplest and at the same time rigorous and most general manner". The fundamental idea of their algebraic-arithmetic approach was to carry over to algebraic function fields the ideas that Dedekind had earlier introduced for algebraic number fields.

Just as an algebraic number field is a finite extension $Q(\alpha)$ of the field Q of rationals, so an algebraic function field is a finite extension $K = \mathbb{C}(z)(w)$ of the field $\mathbb{C}(z)$ of rational functions (in the indeterminate z). That is, w is a root of a polynomial $a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_n\alpha^n$, where $a_i \in \mathbb{C}(z)$ (we can take $a_i \in \mathbb{C}[z]$). Thus $w = f(z)$ is an algebraic function defined implicitly by the polynomial equation $P(z, w) = a_0 + a_1 w + a_2 w^2 + \cdots + a_n w^n = 0$. In fact, all elements of $K = \mathbb{C}(z, w)$ are algebraic functions.

Now let A be the "ring of integers" of K over $\mathbb{C}(z)$; that is, A consists of the elements of K that are roots of *monic* polynomials over $\mathbb{C}[z]$. As for algebraic numbers, here too every nonzero ideal of A is a unique product of prime ideals. Incidentally, in the case of the field of meromorphic functions on a Riemann surface, the role of the integers is played by the entire functions.

Dedekind and Weber were now ready to give a rigorous, algebraic definition of the Riemann surface S of the algebraic function field K: it is (in our terminology) the set of nontrivial discrete valuations on K. The finite points of S correspond to ideals of A; to deal with points at infinity of S Dedekind and Weber introduced the notions of "place" and "divisor". Many of Riemann's ideas about algebraic functions were here developed algebraically and rigorously. In particular, a rigorous proof was given of the important Riemann-Roch theorem. See [3].

Beyond Dedekind and Weber's technical achievements in putting major parts of Riemann's algebraic function theory on solid ground, their conceptual breakthrough lay in pointing to the strong analogy between algebraic number fields and algebraic function fields, hence between algebraic number theory and algebraic geometry. This analogy proved extremely fruitful for both theories. For example, the use of power series in algebraic geometry inspired Hensel in 1897 to introduce p-adic numbers ("power series" in the prime p). The resulting idea of p-adic completion proved important in both algebraic number theory and algebraic geometry. Another noteworthy aspect of Dedekind and Weber's work was its generality and applicability to arbitrary fields, in particular Q and Z_p, which were important in number-theoretic contexts. Thus, ideas from algebraic geometry could be applied to number theory.

(ii) Polynomial rings and their ideals. Polynomial ideals in algebraic geometry had their implicit beginnings in M. Noether's work in the 1870s. Important advances were made by Kronecker in the 1880s and especially by Lasker and Macauley in 1905 and 1913, respectively.

The need for polynomial ideals in the study of algebraic varieties is manifest. An algebraic variety V is defined as the set of points in $\mathbb{R}^n$ (or $\mathbb{C}^n$) satisfying a system of polynomial equations $f_i(x_1, \dots, x_n) = 0$, $i = 1, 2, 3, \dots$. The Hilbert Basis Theorem implies that finitely many equations will do. But different systems of polynomial equations may give rise to the same set of roots. For example, the circle V in $\mathbb{R}^3$ of radius 2 lying in the plane parallel to the (x, y) plane and two units above it may be described as $V = \{(x, y, z)\colon\ x^2 + y^2 - 4 = 0,\ z - 2 = 0\}$, as $V = \{(x, y, z)\colon\ x^2 + y^2 + z^2 - 8 = 0,\ z - 2 = 0\}$, or as $V = \{(x, y, z)\colon\ x^2 + y^2 - 4 = 0,\ x^2 + y^2 - 2z = 0\}$. Is there a canonical set of polynomials that describes the variety (circle) V?

It is easy to see that if $f_1, \dots, f_m$ are polynomials that vanish on the points of V, then so do all polynomials of the set $I = \{g_1 f_1 + \cdots + g_m f_m\colon\ g_i \in \mathbb{R}[x, y, z]\}$. But I is an ideal of the polynomial ring $\mathbb{R}[x, y, z]$. In fact, the set of *all* polynomials of $\mathbb{R}[x, y, z]$ that vanish on the points of V is also an ideal—and it is evidently the "canonical" set of polynomials to describe V.

Note that the preceding remarks point to a correspondence between ideals of $\mathbb{R}[x_1, \dots, x_n]$ (or of $\mathbb{C}[x_1, \dots, x_n]$) and varieties in $\mathbb{R}^n$ (or $\mathbb{C}^n$): If V is a variety, let $I(V) = \{f(x_1, \dots, x_n) \in \mathbb{R}[x_1, \dots, x_n]\colon\ f(a_1, \dots, a_n) = 0 \text{ for all } (a_1, \dots, a_n) \in V\}$, and if J is an ideal of $\mathbb{R}[x_1, \dots, x_n]$, let $V(J) = \{(b_1, \dots, b_n) \in \mathbb{R}^n\colon\ g(b_1, \dots, b_n) = 0 \text{ for all } g \in J\}$. The Hilbert Nullstellensatz, in one of its incarnations, says that $V(J) \neq 0$ if the variety is in $\mathbb{C}^n$, or K^n for any algebraically closed field K. This correspondence is central in algebraic geometry. It is, in fact, a one-to-one correspondence between varieties (over an algebraically closed field K) and their largest defining ideals (the so-called radical ideals). Under this correspondence, prime ideals correspond to irreducible varieties (those that cannot be non-trivially decomposed into finite unions of other varieties). See [3].

Lasker and Macauley exploited this correspondence in the early 20th century by undertaking a thorough study of ideals in polynomial rings in order to shed light on algebraic varieties. Lasker's major result was the "primary decomposition" of ideals: Every ideal in a polynomial ring $F[x_1, \dots, x_n]$ is a finite intersection of primary ideals. Primary ideals, first defined by Lasker, are generalizations of prime ideals; the former are to the latter what prime powers are to primes in the ring of integers. Translated into the language of algebraic geometry, the result says that every variety is a finite union of irreducible varieties. Macauley proved the uniqueness of the primary decomposition, which implied that every variety can be expressed uniquely as a union of irreducible varieties—a type of fundamental theorem of arithmetic for varieties. By the way, it is no easy matter to determine *geometrically* when a curve is irreducible; it is the algebra that comes to the geometer's aid here.

C. THE ABSTRACT DEFINITION OF A RING. In the first decade of the 20th century there were well-established, flourishing, concrete theories of both commutative and noncommutative rings and their ideals. Their roots were in algebraic number theory, algebraic geometry, and the theory of hypercomplex number systems. Moreover, abstract (axiomatic) definitions of groups, fields, and vector spaces had then been in existence for about two decades. The time was ripe for the abstract ring concept to emerge.

The first abstract definition of a ring was given by Fraenkel (of set-theory fame) in a 1914 paper entitled "On zero divisors and the decomposition of rings" [4]. He defines a ring as "a system" with two (abstract) operations, to which he gives the names addition and multiplication. Under one of the operations (addition) the system forms a group (he gives its axioms). The second operation (multiplication) is associative and distributes over the first. Two axioms give the closure of the system under the operations, and there is the requirement of an identity in the definition of the ring. Commutativity under addition does *not* appear as an axiom but is proved! So are other elementary properties of a ring such as $a \times 0 = 0$, $a(-b) = (-a)b = -(ab)$, and $(-a)(-b) = ab$.

Fraenkel's work exerted little influence since it was not grounded in the major concrete theories that had earlier been established. Its main significance was that rings now began to be studied as independent, abstract objects, not just as rings of polynomials, as rings of algebraic integers, or as rings (algebras) of hypercomplex numbers.

D. EMMY NOETHER AND EMIL ARTIN. Yet rings of polynomials, rings of algebraic integers, and rings of hypercomplex numbers remained central in ring theory. In the hands of the master algebraists Noether and Artin their study was transformed in the 1920s into powerful, abstract theories. Noether's two seminal papers of 1921 and 1927 extended and abstracted the decomposition theories of polynomial rings on the one hand and of the rings of integers of algebraic number fields and algebraic function fields on the other, to abstract commutative rings with the ascending chain condition—now called *noetherian rings*.

More specifically, Noether showed in her 1921 paper, "Ideal theory in rings", that the results of Hilbert, Lasker, and Macauley on primary decomposition in polynomial rings hold for any (abstract) ring with the ascending chain condition. Thus results which seemed inextricably connected with the properties of polynomial rings were shown to follow from a single axiom! In her 1927 paper, "Abstract development of ideal theory in algebraic number fields and function fields", she characterized abstract commutative rings in which every nonzero ideal is a unique product of prime ideals. These are now called *dedekind domains*.

Artin, inspired by Noether's work on commutative rings with the ascending chain condition, generalized Wedderburn's structure theorems in his 1927 paper, "On the theory of hypercomplex numbers", to noncommutative semi-simple rings with the descending chain condition. In particular, he showed that such rings (now called *artinian rings*) can be decomposed into direct sums of simple rings that, in turn, are matrix rings over division rings.

While Fraenkel gave the first abstract definitions of a ring, Noether and Artin made the abstract ring concept central in algebra by framing in an abstract setting the theorems that were its major inspirations. In this context they introduced, and gave prominence to, such fundamental algebraic notions as ideal (including one-sided ideal), module, and chain conditions (both ascending and descending). Ring theory now took its rightful place along the by then well-established theories of groups and fields as one of the pillars of abstract algebra.

REFERENCES

1. Adams, W. W., and Goldstein, L. J., *Introduction to Number Theory*, Prentice-Hall, 1976.
2. Edwards, H. M., "The genesis of ideal theory", *Arch. Hist. Ex. Sci.* 23 (1980), 321–378.
3. Eisenbud, D., *Commutative Algebra, with a View Toward Algebraic Geometry*, Springer-Verlag, 1995.

4. Fraenkel, A., "Über die Teiler der Null und die Zerlegung von Ringen", *Jour. für die Reine und Angew. Math.* 145 (1914), 139–176.
5. Ireland, K., and Rosen, M., *A Classical Introduction to Modern Number Theory*, 2nd ed., Springer-Verlag, 1982.
6. van der Waerden, B. L., *A History of Algebra*, Springer-Verlag, 1985.

Department of Mathematics & Statistics
York University
North York, Ontario
CANADA M3J 1P3
kleiner@yorku.ca

A method to trisect a series of angles having relation to each other; also another to trisect any given angle. By James Sabben. 1848 (two quarto pages). "The consequence of years of intense thought."

De Morgan's comment on this effort: Very likely, and very sad.

A. De Morgan, *A Budget of Paradoxes*, Vol. 2, p. 10, Open Court Edition, 1915.

Contributed by M. Bruckheimer, The Weizmann Institute of Science

THE AUTHORS

RON PERLINE received his B.A. from U.C. Santa Cruz, and his Ph.D. at U.C. Berkeley. He has since taught at U.C. Santa Cruz and Drexel University. His research interests include soliton theory, geometry, and various aspects of applied mathematics and numerical analysis. During academic year 95–96 he will be on sabbatical leave at AT & T Bell Labs, helping to shape "light into solitons" with the analysis of light propagation in optical fibers. He enjoys outdoor sports, including running, bicycling and roller blading. Wind, water, and time permitting, he is an avid (although not particularly accomplished) windsurfer.

SABER ELAYDI was born in Beersheba, Palestine from a Bedouin family. At the age of five he moved to a refugee camp in Gaza where he received his pre-college education and developed his life-time love of mathematics. He earned his Ph.D. degree in 1978 from the University of Missouri-Columbia. Since then he held academic positions at Kuwait University, Case Western Reserve University, and the University of Colorado. Currently he is a Professor and Chairman of the Department of Mathematics at Trinity University in San Antonio, Texas. His mathematical interests include difference and differential equations, and topological dynamics. He serves as Co-Editor-in-Chief of the *Journal of Difference Equations and Applications*. Saber Elaydi is the author of the book *An Introduction to Difference Equations and Applications*. His main hobbies are jogging and playing ping-pong.

KIMMO ERIKSSON, born 1967, received his Ph.D. in combinatorics under Anders Björner in 1993 from KTH, Stockholm. Concerned about the general acceptance of boring and unintelligible talks in mathematics, Kimmo does his best to make a difference. He tackled the polygon-splitting problem when he wanted a new nice topic for a scheduled talk. In his free time he enjoys playing chamber music with three other mathematicians, namely his brother, mother and father!

JOHN HANNAH was educated in England, Australia and New Zealand. He, in turn, has taught at universities in Melbourne (Australia), Dublin and Galway (Ireland) and Christchurch (New Zealand). His research articles have all been about aspects of ring theory or linear algebra, but he is also interested in mathematics education and helps run an enrichment program for interested high school students. In his spare time, John coaxes fruit and vegetables from a modest suburban garden.

JOHN SYNOWIEC received his B.S. and M.S. from DePaul University and his Ph.D. in 1964 from Illinois Institute of Technology on pseudo-conformal geometry under John DeCicco. His current interest is in the history of mathematics, especially of harmonic analysis and of distribution theory.

NOTES

(**10528**) These two problems are only loosely connected. In particular, the additional condition $\sum a_i = \sum 1/a_i$ is *not* to be assumed in (**b**). Part (**b**) has been proved for $n = 3$ and $n = 4$.

SOLUTIONS

Recovering Individuals from the Joint Distribution of Sums

10304 [1993, 401]. *Proposed by Ignacy I. Kotlarski, Oklahoma State University, Stillwater, OK.*

Let $\lambda_0, \lambda_1, \lambda_2$ be three positive constants. Let X_0, X_1, X_2 be three independent discrete random variables with nonnegative integer values only. Suppose that $EX_0 = \lambda_0$. Now let $Y_1 = X_0 + X_1$ and $Y_2 = X_0 + X_2$ and suppose that the joint probability distribution for (Y_1, Y_2) is given by

$$P(Y_1 = j_1, Y_2 = j_2) = \sum_{k=0}^{\min(j_1,j_2)} \frac{\lambda_0^k \lambda_1^{j_1-k} \lambda_2^{j_2-k}}{k!\,(j_1-k)!\,(j_2-k)!} e^{-(\lambda_0+\lambda_1+\lambda_2)}$$

for nonnegative integers j_1 and j_2. Find the distributions of X_0, X_1, and X_2.

Solution I by David Callan, University of Wisconsin, Madison, WI. The assumption that $EX_0 = \lambda_0$ is unnecessary. By hypothesis, we have

$$P(Y_1 = a, Y_2 = 0) = \frac{\lambda_1^a}{a!} e^{-(\lambda_0+\lambda_1+\lambda_2)}.$$

But $(Y_1, Y_2) = (a, 0)$ entails $X_0 = X_2 = 0$ and $X_1 = a$; hence by independence, we also have $P(Y_1 = a, Y_2 = 0) = P(X_0 = 0)P(X_1 = a)P(X_2 = 0)$. Comparing these expressions for all values of a yields

$$P(X_1 = a) = K\frac{\lambda_1^a}{a!}$$

for some constant K, forcing X_1 to be Poisson with parameter λ_1. By symmetry, X_2 is Poisson with parameter λ_2. Summing the hypothesized equation over j_2 yields

$$P(X_0 + X_1 = j_1) = \sum_{k=0}^{j_1} \frac{\lambda_0^k}{k!} e^{-\lambda_0} \frac{\lambda_1^{j_1-k}}{(j_1-k)!} e^{-\lambda_1}. \tag{1}$$

However, denoting $P(X_0 = k)$ by p_k and again using independence, we have

$$P(X_0 + X_1 = j_1) = \sum_{k=0}^{j_1} p_k \frac{\lambda_1^{j_1-k}}{(j_1-k)!} e^{-\lambda_1} \tag{2}$$

for all j_1. It is a simple matter to infer inductively from the equality of the right sides of (1) and (2) that $p_k = \lambda_0^k e^{-\lambda_0}/k!$. Thus, X_0, X_1 and X_2 are Poisson with parameters λ_0, λ_1 and λ_2 respectively.

Solution II by Donald A. Darling, Newport Beach, CA. Multiply the given expression by $t_1^{j_1}t_2^{j_2}$ and sum on $j_1, j_2 = 0, 1, 2, \ldots$. The left side will give the joint generating function for Y_1, Y_2, i.e.

$$E\left(t_1^{Y_1}t_2^{Y_2}\right) = E\left(t_1^{X_0+X_1}t_2^{X_0+X_2}\right) = E\left((t_1t_2)^{X_0}\right)E\left(t_1^{X_1}\right)E\left(t_2^{X_2}\right)$$

when we use the mutual independence of X_0, X_1 and X_2. The right side is easily identified as

$$e^{t_1\lambda_1+t_2\lambda_2+t_1t_2\lambda_0-\lambda_0-\lambda_1-\lambda_2} = e^{\lambda_1(t_1-1)}e^{\lambda_2(t_2-1)}e^{\lambda_0(t_1t_2-1)}.$$

When $t = 0$, the generating function $E\left(t^X\right)$ is $P(X = 0)$; thus, we may set $t_2 = 0$ to obtain

$$E\left(t_1^{X_1}\right)P(X_2 = 0)P(X_0 = 0) = e^{\lambda_1(t_1-1)}e^{-\lambda_2-\lambda_0}.$$

Dividing this by the same equation with $t_1 = 1$, we obtain $E\left(t_1^{X_1}\right) = e^{\lambda_1(t_1-1)}$. Similarly, $E\left(t_2^{X_2}\right) = e^{\lambda_2(t_2-1)}$. Finally, $E\left((t_1t_2)^{X_0}\right) = e^{\lambda_0(t_1t_2-1)}$. These formulas show that each X_i has Poisson distribution with mean λ_i, $i = 0, 1, 2$. It appears unnecessary to require *ab initio* that X_0 has mean λ_0 or, for that matter, that the random variables X_0, X_1, X_2 be integer valued.

Editorial comment. The proposer's solution used generating functions as in Solution II. However, the distribution of X_0 was attacked first, and this appeared to require the assumption that $EX_0 = \lambda_0$. In contrast to this cautious approach, two readers found the factored form of the generating function and immediately concluded that this required that the factors must be the generating functions of the individual variables. These solutions were judged incomplete. Two other solvers noted that such a result is a theorem of Raikov (see E. Lukacs, *Characteristic Functions*, Charles Griffin, London, 1960, p. 243 or M. Loeve, *Probability*, Vol. I, fourth edition, Springer-Verlag, 1977 for a precise statement and proof of Raikov's theorem).

Gérard Letac generalized this result in two different directions.

First, he considered certain larger collections of sums of the form $Y_i = \sum_{j=0}^{n} a_{ij}X_j$, for $i = 1, \ldots, k$ with $a_{ij} \in \{0, 1\}$. The admissible collections are those for which the vectors $(a_{i0}a_{i'0}, \ldots, a_{in}a_{i'n})$, where i and i' run through $1, \ldots, k$ independently, span $\mathbb{R}^{n+1}$.

In addition, the Poisson distribution to be obtained as the distribution of the X_j is replaced by a one parameter family of distributions μ_λ on $[0, \infty)$, with $\lambda > 0$, such that the Laplace transform of μ_λ is equal to the Laplace transform of μ_1 raised to the power λ.

Assuming that X_j has distribution μ_{λ_j}, a joint distribution of the Y_i is determined. The result is that this distribution of the Y_i forces the X_j to have distribution μ_{λ_j}.

Solved also by R. J. Chapman (U. K.), P. J. Fitzsimmons, J. R. Hoffman, I. Kastanas, S. C. Kian (Singapore), G. Letac (France), O. P. Lossers (The Netherlands), K. Wagner, A. N. 't Woord (The Netherlands), GCHQ Problem Solving Group (U. K.), Western Maryland College Problems group, and the proposer. Two incomplete solutions were received.

Canonical Examples of Groups

10307 [1993, 498]. *Proposed by John Calvin Williams, student, and I. Martin Isaacs, University of Wisconsin, Madison WI.*

Can one construct a set $\mathcal{X}$ of finite groups satisfying the two conditions:

i. $\mathcal{X}$ contains precisely one representative from each isomorphism class.

ii. If $A \in \mathcal{X}$ is isomorphic to a subgroup of $B \in \mathcal{X}$, then A *is* a subgroup of B.

Solution by Reiner Martin, Deutsche Bank, Frankfurt, Germany. No. Otherwise, let S_n and C_n be elements of $\mathcal{X}$ isomorphic to the symmetric group on n elements and to the cyclic

group of order n, respectively. If S_4 acts as a permutation group on four letters, it is easy to see that S_3 must be the stabilizer of some letter and that C_4 is generated by a 4-cycle. Thus $S_3 \cap C_4$ is trivial. This is a contradiction, since both contain C_2.

Editorial comment. Vladimir Božin and Robin J. Chapman each noted that the answer would be *Yes* if $\mathcal{X}$ were restricted to finite abelian groups.

Solved also by D. Alvis, R. Barbara (Lebanon), V. Božin (student, Yugoslavia), R. J. Chapman (U. K.), S. M. Gagola Jr., R. Holzsager, O. P. Lossers (The Netherlands), the MMRS group of Oklahoma State University, and the proposers.

A Collinear Configuration

10308 [1993, 498]. *Proposed by Robert Connelly and John H. Hubbard, Cornell University, Ithaca, NY, and Walter Whiteley, York University, North York, Ontario, Canada.*

Suppose that $p_1, p_2, p_3, q_1, q_2, q_3$ are six points in the plane and that the distance between p_i and q_j $(i, j = 1, 2, 3)$ is $i + j$. Show that the six points are collinear.

Solution I by Ilias Kastanas, California State University, Los Angeles, CA. Let the coordinates of p_i, q_j be (x_i, y_i), (a_j, b_j) respectively. Without loss of generality, let $(x_2, y_2) = (2, 0)$ and $(a_2, b_2) = (-2, 0)$. Then we have the equations

$$(x_i - a_j)^2 + (y_i - b_j)^2 = (i + j)^2. \qquad (E_{ij})$$

By taking $E_{11} - (E_{12} + E_{21})$, $E_{33} - (E_{32} + E_{23})$, $E_{13} - (E_{12} + E_{23})$, and $E_{31} - (E_{32} + E_{21})$ we get

$$(x_1 - 2)(a_1 + 2) + y_1 b_1 + 1 = 0,$$
$$(x_3 - 2)(a_3 + 2) + y_3 b_3 + 1 = 0,$$
$$(x_1 - 2)(a_3 + 2) + y_1 b_3 - 1 = 0,$$
$$(x_3 - 2)(a_1 + 2) + y_3 b_1 - 1 = 0.$$

Therefore, $(1 + y_1 b_1)(1 + y_3 b_3) = (x_1 - 2)(a_1 + 2)(x_3 - 2)(a_3 + 2) = (1 - y_1 b_3)(1 - y_3 b_1)$, from which $(y_1 + y_3)(b_1 + b_3) = 0$. Suppose that $y_1 + y_3 = 0$ (the case $b_1 + b_3 = 0$ is similar). Then, adding the first and last of the four equation displayed above, we get $(a_1 + 2)(x_1 + x_3 - 4) = 0$. If $a_1 + 2 = 0$, E_{21} would give the contradiction $b_1^2 = -7$, so $x_1 + x_3 = 4$. By $E_{32} - E_{12}$, it follows that $8(x_3 - x_1) = 16$, so $x_1 = 1$, $x_3 = 3$, and E_{12}, E_{32} then imply $y_1 = y_3 = 0$. It follows that $a_1 = -1$, $a_3 = -3$, $b_1 = b_3 = 0$, and all six points are on the x-axis.

Solution II by MMRS, Oklahoma State University, Stillwater, OK. We shall view the points as complex numbers. Consider four points p_i, p_j, q_k, and q_l, with $i \neq j$ and $k \neq l$. Suppose the line segments $p_i q_k$ and $p_j q_l$ have a point x in common (which might be an endpoint). Then, by the triangle inequality, we have

$$|x - p_i| + |x - q_l| \geq |p_i - q_l|$$

and

$$\left|x - p_j\right| + |x - q_k| \geq \left|p_j - q_k\right|;$$

where equality holds in both inequalities if and only if p_i, p_j, q_k, q_l are all collinear and x is on the line segments $p_i q_l$ and $p_j q_k$. Adding the inequalities, we obtain

$$|x - p_i| + |x - q_l| + \left|x - p_j\right| + |x - q_k| \geq |p_i - q_l| + \left|p_j - q_k\right|,$$

and, since x is on both $p_i q_k$ and $p_j q_l$, this simplifies to

$$|p_i - q_k| + \left|p_j - q_l\right| \geq |p_i - q_l| + \left|p_j - q_k\right|.$$

However, by hypothesis both sides are equal to $i + j + k + l$. Therefore, equality does hold in all inequalities above, and so p_i, p_j, q_k, q_l are collinear. Furthermore, since x is on all four segments connecting p's to q's, these four points must be ordered so that the p's are on the opposite side of x from the q's.

Now, the line segments connecting all three p's to all three q's form a complete bipartite graph $K_{3,3}$. Since this graph is well known to be nonplanar, there must be two segments p_iq_k and p_jq_l with $i \neq j$ and $k \neq l$ that have a point x in common. By the above, the four points p_i, p_j, q_k, q_l are then collinear. We may also assume that they appear on their common line in this order. Consider the remaining two points p_m and q_n. Apply the above argument to p_i, p_m, q_l, and q_k with $x = q_k$, we see that p_m is also on the line containing p_i, p_j, q_k, and q_l. Similarly, q_n is on this line.

Editorial comment. Most solvers used coordinates of the points, sometimes encoded as vectors or complex numbers, leading to a system of equations resembling that considered in Solution I. Geometric considerations often guided the solution of these equations. Another approach was taken by Frank Schmidt. Nine of the fifteen distances between the six points are given, and he introduced six variables representing the remaining distances. Using the vanishing of the Cayley-Menger determinant (see M. Berger, *Geometry I*, Springer-Verlag, 1977, p. 239) as a criterion for four points to lie in a plane, he obtained equations relating them . These equations have a unique solution. These distances can then be used in general to locate the points in the plane; and in this case, to verify that they lie on a line.

Solved also by E. Aichinger (student, Austria), D. Alvis, J. Anglesio (France), R. Barbara (Lebanon), V. Božin (student, Yugoslavia), M. Brahm, R. J. Chapman (U. K.), H. S. Gunaratne (Brunei), J. G. Heuver (Canada), R. D. Hurwitz, O. P. Lossers (The Netherlands), A. D. Melas (Greece), I. Praton & E. P. Venugopal (student), P. Rennert, F. Schmidt, N. S. Thornber, PCC Math Problem Solvers Group, and the proposers.

Binomial Coefficient Growth

10310 [1993, 499]. *Proposed by E. Rodney Canfield, University of Georgia, Athens, GA.*

Fix an integer $r \geq 2$. Using Stirling's formula we may find constants c_1 and c_2 such that

$$\binom{rm}{m} \sim \frac{c_1(c_2)^m}{m^{1/2}}$$

as $m \to \infty$. Prove that the ratio $\binom{rm}{m} m^{1/2}/c_2^m$ is an increasing function of m for $m \geq 1$.

Solution by MMRS, Oklahoma State University, Stillwater, OK. Taking logarithms, we see that it suffices to prove that

$$f(m) = \log\Gamma\big(rm+1\big) - \log\Gamma\big(m+1\big) - \log\Gamma\big((r-1)m+1\big) + \frac{1}{2}\log m - m\log c_2$$

is increasing in m when $m \geq 1$ and $r \geq 2$. By Binet's formula for $\log\Gamma(z)$ (see E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Fourth Edition, Cambridge, 1927, p. 249),

$$\log\Gamma(z+1) = \log z + \log\Gamma(z) = g(z) + h(z)$$

where

$$g(z) = \left(z + \frac{1}{2}\right)\log z - z + C,$$

with $C = (1/2)\log 2\pi$, and

$$h(z) = \int_0^\infty j(t)e^{-tz}\,\frac{dt}{t}$$

with

$$j(z) = \frac{1}{2} - \frac{1}{t} + \frac{1}{e^t - 1}$$

By a straightforward calculation with $c_2 = r^r \,/\, (r-1)^{r-1}$:

$$g\big(rm\big) - g\big((r-1)m\big) - g\big(m\big) + \frac{1}{2}\log m - m\log c_2 = \frac{1}{2}\log\frac{r}{r-1} - C.$$

Thus, it suffices to show that

$$\int_0^\infty j(t)\Big(e^{-rmt} - e^{-(r-1)mt} - e^{-mt}\Big)\frac{dt}{t}$$

is an increasing function of m. First, note that $j(t) \geq 0$ for $t \geq 0$. This is equivalent to $n(t) = 2t(e^t-1)j(t) = (t-2)e^t + t + 2$ being nonnegative. However, $n'(t) = (t-1)e^t + 1$ and $n''(t) = te^t$. Hence $n(0) = n'(0) = 0$ and $n''(t) = te^t \geq 0$ for $t \geq 0$, from which the nonnegativity of $n(t)$ follows. It now comes down to showing that $e^{-rmt} - e^{-(r-1)mt} - e^{-mt}$ is an increasing function of m for each $t > 0$. This expression is $k\big(e^{-tm}\big)$ where

$$k(u) = u^r - u^{r-1} - u.$$

Now

$$k'(u) = ru^{r-1} - (r-1)u^{r-2} - 1 = r(u-1)u^{r-2} + u^{r-2} - 1 > 0$$

if $u > 1$ and $r \geq 2$.

Editorial comment. Note that r need not be assumed to be an integer, and that m is also a real variable in the selected proof. While Jean Anglesio and David M. Bloom also worked in this level of generality, the other solvers restricted attention to integer values of m and r. Typically, this allowed the use of classical inequalities. The proposer's solution, while similar to the selected solution, studied $f(m+1) - f(m)$, indicating a restriction to integer m.

Solved also by D. Alvis, J. Anglesio (France), D. M. Bloom, V. Božin (student, Yugoslavia), S. Byrd & T. J. Walters, M. Carlehed (Sweden), R. J. Chapman (U. K.), D. A. Darling, I. Kastanas, K. S. Kedlaya (student), J. Marengo, A. D. Melas (Greece), R. M. Robinson, H.-J. Seiffert (Germany), and the proposer. Two incorrect or incomplete solutions were received.

A Property that Limits Size

10318 [1993, 590]. *Proposed by William P. Wardlaw, United States Naval Academy, Annapolis MD.*

Suppose that A is an n by n matrix with rational entries whose multiplicative order is 15; i.e. $A^{15} = I$, an identity matrix, but $A^k \neq I$ for $0 < k < 15$. For which n can one conclude from this that

$$I + A + \cdots + A^{14} = 0?$$

Solution by Stephen M. Gagola, Jr., Kent State University, Kent, OH. The answer is $n \leq 6$ (although for $n < 6$ there is no matrix of order 15, so the statement is only vacuously true). We prove a more general result. If m is an integer greater than one with prime power factorization $m = p_1^{e_1} \cdot p_2^{e_2} \cdot \cdots \cdot p_r^{e_r}$, define

$$\begin{aligned}\psi(m) &= \phi(p_1^{e_1}) + \phi(p_2^{e_2}) + \cdots + \phi(p_r^{e_r}) \\ &= p_1^{e_1-1}(p_1 - 1) + p_2^{e_2-1}(p_2 - 1) + \cdots + p_r^{e_r-1}(p_r - 1).\end{aligned}$$

and set

$$\psi_0(m) = \begin{cases} \psi(m), & \text{if } m \text{ is divisible by 4, or if } m \text{ is odd} \\ \psi(m_0), & \text{if } m = 2m_0, \text{ where } m_0 \text{ is odd and greater than 1} \\ 1, & \text{if } m = 2. \end{cases}$$

Letting $n = \psi_0(m)$, we prove that

1. If $k < n$, there does not exist a $k \times k$ rational matrix of order m.
2. If A is an $n \times n$ rational matrix of order m, then $I + A + \cdots + A^{m-1} = 0$.
3. If $k > n$, there exists a $k \times k$ rational matrix of order m with $I + A + \cdots + A^{m-1} \neq 0$.

Suppose n_0 is the smallest integer for which there is a rational $n_0 \times n_0$ matrix B of order m. For $n > n_0$, forming a block diagonal matrix using B and an identity matrix of size $n - n_0$ produces an $n \times n$ matrix of order m such that $I + A + A^2 + \cdots + A^{m-1} \neq 0$.

Moreover, 1 cannot be an eigenvalue of B, as then B itself would be similar to a block diagonal decomposition of an identity matrix with a smaller matrix of order m, contradicting the definition of n_0. Since the minimal polynomial of B has no repeated factors, B is diagonalizable, and it readily follows that $I + B + B^2 + \cdots + B^{m-1} = 0$.

The proof is completed by proving that $n_0 = \psi_0(m)$. The companion matrix of the cyclotomic polynomial $\Phi_{p^e}(x)$ corresponding to a primitive p^e-th root of unity has order p^e and size $\phi(p^e) = p^{e-1}(p-1)$. A block diagonal sum of matrices of this form for every prime power appearing in the factorization of m shows that $n_0 \leq \psi(m)$. Moreover, if $m = 2m_0$, where m_0 is odd, and if A is a matrix of order m_0, then $-A$ has order $2m_0$. Thus $n_0 \leq \psi_0(m)$.

Finally, let A be any $n_0 \times n_0$ matrix of order m. The matrix A is rationally similar to a block diagonal decomposition of irreducible matrices. The characteristic polynomial of any one of these must be an irreducible factor of $x^m - 1$, and hence it has the form $\Phi_d(x)$ for some $d|m$. Without loss of generality, A is in block diagonal form in which the blocks are companion matrices C_d corresponding to cyclotomic polynomials $\Phi_d(x)$ for various $d|m$.

By the definition of n_0, at most one such companion matrix occurs for each divisor d. Moreover, suppose a companion matrix C_d occurs corresponding to a divisor of d that is not a prime power. Write $d = uv$ where $\gcd(u, v) = 1$ and $u, v > 1$. If $\phi(uv) > \phi(u) + \phi(v)$, then C_d may be replaced by the smaller matrix $\begin{pmatrix} C_u & 0 \\ 0 & C_v \end{pmatrix}$, without affecting the order of A, which contradicts the choice of n_0.

The condition $\phi(uv) > \phi(u) + \phi(v)$ holds except when u or v is 2 or when $\{u, v\} = \{3, 4\}$. Neither case arises when m is odd. If m is divisible by 4 and the first case occurs, then some companion matrix C_d for $4|d$ must occur in the decomposition of A. But this implies that a matrix of the form C_2 may be dropped from the decomposition of A without affecting the order of A (contradicting the choice of n_0). If $d = 12$, then there is no contradiction, but the 4×4 matrix C_{12} may be replaced by the 4×4 matrix $\begin{pmatrix} C_3 & 0 \\ 0 & C_4 \end{pmatrix}$, without affecting the order of A.

Thus we may assume that all the companion matrices C_d appearing in a decomposition of A correspond to the divisors d of m which are prime powers or twice an odd prime power. Further, the second possibility can occur only when m is twice an odd number. Since the order of A is exactly m, every prime power occurring in the factorization of m must divide at least one of these d's (for which a corresponding companion matrix C_d appears in A), and the reverse inequality $n_0 \geq \psi_0(m)$ follows.

Editorial comment. Generalizations equivalent to the one above were given by Roy Barbara, Kevin Brown &Daniel Dufresne, Charles Lanski, F. J. Flanigan, Jonathan Merzel, Geoffrey R. Robinson, and A. N. 't Woord. F. J. Flanigan and Frank Schmidt each noted that the problem of finding the smallest possible value of n for which there exists an $n \times n$ rational matrix of order m appears as Problem 445 in *The College Mathematics Journal* [1991, 71; 1992, 74].

Solved also by D. Alvis, R. Barbara (Lebanon), F. R. Beyl, W. Blumberg, K. Brown & D. Dufresne, R. J. Chapman (U. K.), F. J. Flanigan, T. H. Foregger, H. S. Gunaratne (Brunei), R. Holzsager, N. Jensen (Germany), I. Kastanas, K. S. Kedlaya (student), H. Kiechle (Germany), N. Komanda, D. W. Koster, J. Kuplinski, J. F. Kurtzke, C. Lanski, C.-K. Li, J. H. Lindsey II, O. P. Lossers (The Netherlands), L. E. Mattics, J. Merzel, G. R. Robinson, F. Schmidt, A. N.

't Woord (The Netherlands), Anchorage Math Solutions Group, the MMRS group of Oklahoma State University, Western Maryland College Problems group, and the proposer. Five incorrect solutions were received.

Collaborating editors: *David F. Appleyard, Paul T. Bateman, Duane M. Broline, Ezra A. Brown, Barry W. Brunson, Frank S. Cater, Gulbank D. Chakerian, Underwood Dudley, Gerald A. Edgar, Michael A. Filaseta, Ira M. Gessel, Richard A. Gibbs, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Mourad E. H. Ismail, Murray Klamkin, Daniel J. Kleitman, Frederick W. Luttmann, Frank B. Miles, Richard Pfiefer, Stephen L. Portnoy, J. O. Shallit, John Henry Steelman, Kenneth B. Stolarsky, David E. Tepper, Douglas B. Tyler, Daniel Ullman, Charles L Vanden Eynden, William E. Watkins, and Gregory P. Wene.*

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington, IN 47405

A Guide to Distribution Theory and Fourier Transforms. By Robert Strichartz. CRC Press, Boca Raton, 1994, 213 pp.

Reviewed by **John A. Synowiec**

"Distribution? You mean probability distribution.—No, a Schwartz distribution.—Oh... if you are interested in that sort of thing you'll have to talk to somebody else. I don't have much use for them in my work."

This was a conversation with several professors during my days as a graduate student. It reflected an attitude common to many analysts, some of whom later came to regret this condescension towards generalized functions. A review by Freeman Dyson [5] points out another common view at the time. Noting that it is currently fashionable to use distribution theory in studies of quantum field theory, and that reference to Schwartz's *Théorie des distributions* is a must, Dyson went on to say that some authors not only quoted the book but even showed evidence of having read it. Since then, the situation has changed greatly; distributions (and other generalized functions) are now used routinely by mathematicians. However, distribution theory is still not part of the standard undergraduate curriculum in mathematics.

The book under review is a textbook for undergraduates, and as its title indicates, it is not limited to distribution theory. In fact, the book deals with an interconnected triad: generalized functions, Fourier analysis, and partial differential equations. In recent years, undergraduate courses in Fourier analysis have become more popular, as reflected by the appearance of numerous fine textbooks, e.g., the books of Folland [6], Körner [14], J. Walker [28], and P. Walker [29]. Of course, all of these devote some space to partial differential equations, and Folland has a fine chapter on generalized functions, but the primary emphasis is on Fourier series and integrals.

On the other hand, several good textbooks providing an elementary approach to distribution theory have also appeared recently, e.g., Friedlander [7], Kanwal [12], Richards and Youn [18]; and a list of recent books on partial differential equations would include Bleecker and Csordas [2], Gustafson [9], and Strauss [25], among others.

However, Strichartz's blend of generalized functions, Fourier analysis, and partial differential equations is unique. It precisely this integrated basic triad that is most valuable, and one can argue, should be presented to all undergraduates majoring in mathematics. It is beautiful, elegant mathematics, and it is quite useful as well.

Historically, the connections between partial differential equations and Fourier analysis, and even generalized functions, can be traced to the earliest days of these subjects, in the work of d'Alembert and Euler. (See Lützen [16] or Truesdell

[26, 27].) At the time, the concept of function was not clearly established, so it may seem strange to talk about generalized functions; however, Euler did use "generalized" solutions of partial differential equations, which did not satisfy the equations in the normal sense. Fourier series and integrals were developed by Fourier, Cauchy, and Poisson to solve partial differential equations arising in heat conduction, water waves, and other physical phenomena. They often made use of reasoning that, with hindsight, we may describe as involving delta-functions.

As Strichartz points out in his preface, Lebesgue's theory of integration was a major revolution in 20th century analysis. He regards it as one of two such revolutions, but this seems too modest a number. It had great impact, almost immediately, in Fourier analysis: the theories of Fourier series and Fourier integrals were entirely recast in the framework of Lebesgue spaces, especially L^2. Eventually, Lebesgue's theory led to generalized solutions of partial differential equations, which belong to "Sobolev spaces". These involve weak derivatives of functions, which are based on integration by parts. Thus, a function f in some Lebesgue space has weak derivative g with respect to x if g is a function satisfying

$$\int f\frac{\partial\varphi}{\partial x} = -\int g\varphi, \quad \text{for all functions } \varphi \text{ of class } C^\infty \text{ having compact support.}$$

Since their introduction in the 1920s and 1930s, Sobolev spaces have been very useful in partial differential equations. However, most of this work had little or no connection with Fourier analysis. The historical development of the theory of distributions is quite interesting, and involves many parts of mathematics and its applications. A nice presentation of this may be found in Lützen's book [15].

It was Laurent Schwartz, who, in 1945, [21, 22] launched another major revolution in 20th century analysis with his theory of distributions. (This is Strichartz's second one.) His immediate stimulus was from partial differential equations: he did not want to consider generalized solutions that did not have meaning for each term in the equation. He generalized the Sobolev spaces, as Sobolev himself had done, by considering generalized functions to be continuous linear functionals on spaces $\mathscr{D}$ of C^∞-functions with compact support. These functionals, which he called distributions, include all locally integrable functions f by the rule:

$$f[\varphi] = \int f\varphi, \quad \text{for all } \varphi \text{ in } \mathscr{D}.$$

Guided by this rule and the definition of weak derivative, differentiation is extended to every distribution T:

$$\frac{\partial T}{\partial x}[\varphi] = -T\left[\frac{\partial\varphi}{\partial x}\right], \quad \text{for all } \varphi \text{ in } \mathscr{D}.$$

But Schwartz went beyond Sobolev by developing a harmonic analysis for distributions. For this, the space $\mathscr{D}$ was unsuitable, so he introduced what is now known as the Schwartz space $\mathscr{S}$ of all C^∞-functions φ with rapid decay at infinity:

$$|x^\alpha \partial^\beta\varphi(x)| \to 0 \quad \text{as } |x| \to \infty \text{ for all multi-indices } \alpha \text{ and } \beta.$$

Continuous linear functionals on this space $\mathscr{S}$ are called tempered distributions, and may be considered to be a subclass of the distributions on $\mathscr{D}$. Tempered distributions have become very useful in partial differential equations, and, because many basic operations of analysis are easy and natural with tempered distributions, they provide a very useful framework for Fourier analysis. See, e.g., Khavin [13].

The complete theory appeared in Schwartz's two-volume treatise, *Théorie des distributions*, [**23**], published in 1950 and 1951. This quickly led to important results in partial differential equations. For example, the concept of "fundamental solution" was rather clumsy in the classical theory, but in distribution theory it is easy: a distribution E is a fundamental solution of a partial differential operator P if

$$PE = \delta,$$

where δ is the delta-distribution defined by

$$\delta[\varphi] = \varphi(0), \qquad \text{for all } \varphi.$$

To solve a linear partial differential equation $Pu = f$, where f is a function of class C^∞, it suffices to find a fundamental solution, say a tempered distribution E, of P. In this case, setting $u = E * f$, the convolution of E and f, yields a solution:

$$P(E * f) = (PE) * f = \delta * f = f,$$

by properties of convolution of distributions. Of course, this requires the existence of fundamental solutions, and one of many impressive results obtained using distributions and Fourier analysis was a proof in the mid-1950s of the existence of fundamental solutions by L. Ehrenpreis and by B. Malgrange.

Curiously, it was a negative result that created interest in distributional solutions of partial differential equations: in 1957, H. Lewy showed that the partial differential equation

$$-i\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} - 2(x + iy)\frac{\partial u}{\partial z} = f,$$

has no solutions on any non-empty open in R^3, for certain functions f of class C^∞. This was a most unexpected result! L. Hörmander showed that Lewy's equation has no distributional solution either. The fact that distributions can't provide solutions for all linear partial differential equations apparently convinced many mathematicians that this was not just "abstract nonsense".

However, the real power provided by the union of generalized functions and Fourier analysis for use on partial differential equations was revealed when equations with variable coefficients were attacked. In this case, Fourier transforms were insufficient. In the mid 1960s, Kohn and Nirenberg, Hörmander, and others developed the theory of pseudodifferential operators, which have the form

$$Au(x) = \frac{1}{(2\pi)^n}\int_{R^n} \hat{u}(\xi)a(x,\xi)e^{-ix\cdot\xi}\,d\xi,$$

where $a(x, \xi)$ is a function called the *symbol* of the operator A, and $\hat{u}$ is the Fourier transform of u.

It is ironic that the historical path to pseudodifferential operators did not proceed by a natural way, via Fourier transforms, but rather through another part of harmonic analysis, Calderón and Zygmund's theory of singular integral operators, developed in the 1950s. To be useful, properties of the symbol must be known precisely, and tempered distributions are usually used to overcome problems due to lack of absolute convergence of the integrals. The resulting operational calculus, the theory of pseudodifferential operators, currently plays a major role in the general theory of linear partial differential equations. A detailed introduction to all of this, the triad under discussion, may be found in the massive, four-volume masterpiece on partial differential operators by Hörmander [**10**].

Strichartz (page 175) calls the theory of pseudodifferential operators

> "one of the glorious achievements of mathematical analysis in the last quarter century."

and the author of a recent textbook on pseudodifferential operators says

> "...the development of this theory has reached such a state that the basic results can be considered as a complete whole, and should be mastered by all mathematicians, especially those involved in analysis."
>
> Saint-Raymond [20, page vii]

It is time to describe Strichartz's book. Our triad forms the three main themes of the book, which falls naturally into two parts. Part I consists of the first five chapters, which cover all of the main points: distributions and their basic properties, Fourier transforms of functions of Schwartz's class and of tempered distributions, and partial differential equations. Part II consists of chapters 6, 7, and 8, which fill in more details on the three themes: the structure of distributions, Fourier analysis, and Sobolev spaces and microlocal analysis. Part II is an extension of the basic course, meant to satisfy the curiosity aroused in Part I and as an appetizer for further study of these topics. It is relevant that Hörmander [11] has slightly revised the first volume, *Distribution Theory and Fourier Analysis*, of his treatise, stating in the preface

> "...this volume has been written as a general course in modern analysis on a graduate level and not only as the beginning of a specialized course in partial differential equations." [page iv].

Strichartz's book can serve as an introduction to this work, although a large gap still remains between levels of the two.

Part I is based on lectures and is a clearly-written, well-motivated exposition of the most basic parts of the triad treated. The second part, written for this book, is more abstract, and covers more difficult material, but is also very readable. Some of the topics included in Part II are: Paley-Wiener theorems, the Heisenberg Uncertainty Principle, Haar functions and wavelets, pseudodifferential operators, wave front sets, and microlocal analysis of singularities. Strichartz includes his intriguing *doctrine of microlocal myopia*: pay attention to the singularities, and other issues will take care of themselves. Exercises are provided for each chapter in the book.

The official prerequisites are multidimensional calculus and some complex analysis. The book can be read without the latter, but this would involve omitting several passages involving contour integrals. Lebesgue integration is not assumed.

Strichartz's way of handling the problem of Lebesgue integration in undergraduate courses, e.g., for Fourier analysis, distribution theory, functional analysis, is to advise the reader to interpret any integral in the book in any familiar sense, Lebesgue or Riemann. Of course, L^p-norms appear in places in the book, e.g., in connection with Sobolev spaces. Some authors, e.g., Körner [14], basically ignore Lebesgue integration, and present the subject in terms of Riemann integrals. The opposite approach is that of Debnath and Mikusinski [3], who include a mini-course on Lebesgue integration in their book. An intermediate position is taken by Folland [6], who does not assume prior knowledge, but states the few properties of

Lebesgue integrals used in his book, without any systematic presentation of this theory. This seems to be an admirable choice.

The reason Folland's approach seems best is that it emphasizes the importance of the Lebesgue integral as it appears in action. Books such as Folland's (and Strichartz's) can give students incentive for the study of courses such as "Mathematical Analysis", which often seem dry and boring to them. Instead of seeing these courses as merely "calculus all over again", albeit made rigorous, they can see such courses as presenting material, e.g., Lebesgue integration, that can be useful and interesting. This does not require a full undergraduate course on integration theory. There are many books on "Analysis" that present Lebesgue integration, e.g., Apostol **[1]**, Goldberg **[8]**, Mikusinski and Mikusinski **[17]**, and Rudin **[19]**; there is even a textbook that presents the Riemann-style gauge integral: DePree and Swartz **[4]**. In light of the excellence of Strichartz's book under review, his forthcoming book on analysis **[24]** presents a tantalizing prospect.

To summarize: the material covered by Strichartz's book should probably be part of the undergraduate core of every mathematics student; it involves some beautiful, important mathematics that can stand on its own, and has many applications. Strichartz's presentation of this material is remarkable, and may stimulate the student to further study. It is a marvelous book, a type that mathematics sorely needs.

REFERENCES

1. Apostol, Tom M., *Mathematical Analysis*. Addison-Wesley, 2nd edition, 1974, 492 pp.
2. Bleecker, David, and Csordas, George, *Basic Partial Differential Equations*. Van Nostrand and Reinhold, 1992, 676 pp.
3. Debnath, Lokenath, and Mikusinski, Piotr, *Introduction to Hilbert Spaces With Applications*. Academic Press, 1990, 509 pp.
4. DePree, John D., and Swartz, Charles W., *Introduction to Real Analysis*. John Wiley, 1988, 355 pp.
5. Dyson, F., book review, *Physics Today*, (1958) No. 6, p. 28.
6. Folland, Gerald B., *Fourier Analysis and its Applications*. Wadsworth, 1992, 433 pp.
7. Friedlander, F. G., *Introduction to the Theory of Distributions*. Cambridge University Press, 1982, 157 pp.
8. Goldberg, R. R., *Methods of Real Analysis*. John Wiley, 2nd edition, 1976, 402 pp.
9. Gustafson, Karl, *Introduction to Partial Differential Equations and Hilbert Space Methods*. John Wiley, 2nd edition, 1987, 409 pp.
10. Hörmander, Lars, *The Analysis of Linear Partial Differential Operators*. Springer, Vol. I, 1983, 391 pp; Vol. II, 1983, 391 pp; Vol. III, 1985, 524 pp; Vol. IV, 1985, 352 pp.
11. _____, *The Analysis of Linear Partial Differential Operators*. Springer, Vol. I, 2nd edition, 1990, 440 pp.
12. Kanwal, Ram P., *Generalized Functions: Theory and Technique*. Academic Press, 1983, 428 pp.
13. Khavin, V. P., "Methods and Structure of Commutative Harmonic Analysis", pp. 1–111 of *Commutative Harmonic Analysis* I. Ed. V. P. Khavin and N. K. Nikolskii. Springer, 1991, 268 pp.
14. Körner, Thomas W., *Fourier Analysis*. Cambridge University Press, 1988, 591 pp.
15. Lützen, Jesper, *The Prehistory of the Theory of Distributions*. Springer, 1983, 232 pp.
16. _____, "Euler's vision of a general partial differential calculus for a generalized kind of function", *Math. Magazine* 56, (1983) pp. 299–306.
17. Mikusinski, Jan, and Mikusinski, Piotr, *An Introduction to Analysis*. John Wiley, 1993, 290 pp.
18. Richards, J. Ian, and Youn, Heckyung K., *Theory of Distributions: a Non-Technical Introduction*. Cambridge University Press, 1990, 147 pp.
19. Rudin, Walter, *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976, 342 pp.
20. Saint Raymond, Xavier, *Elementary Introduction to the Theory of Pseudodifferential Operators*. CRC Press, 1991, 108 pp.
21. Schwartz, Laurent, "Généralisation de la notion de fonction, de transformation de Fourier, et applications mathématiques et physiques", *Annales Univ. Grenoble* 21 (1945) pp. 57–74.

22. ______, "Théorie des distributions et transformations de Fourier", *Ann. Univ. Grenoble* 23 (1948) pp. 7–24.
23. ______, *Théorie des distributions*. Hermann & Cie Tome I, 1950, 148 pp; Tome II, 1951, 169 pp.
24. Strichartz, Robert S., *The Way of Analysis*. Jones & Bartlett, 1995, ~ 478 pp.
25. Strauss, Walter, *Partial Differential Equations: An Introduction*. John Wiley, 1992, 425 pp.
26. Truesdell, Clifford, "The influence of elasticity on analysis: the classical heritage", *Bull. Amer. Math. Soc.* 9, (1983) pp. 293–310.
27. ______, *An Idiot's Fugitive Essays on Science*. Springer, 1984, 661 pp.
28. Walker, James S., *Fourier Analysis*. Oxford University Press, 1988, 440 pp.
29. Walker, P. L., *The Theory of Fourier Series and Integrals*. John Wiley, 1986, 192 pp.

Department of Mathematics
Indiana University Northwest
3400 Broadway
Gary, IN 46408

Answer to Picture Puzzles

(pp. 385 and 392)

George Uhlenbeck and Karen Uhlenbeck (Karen Keskula), who was once married to Olke Uhlenbeck, George's son. See the April, 1996 issue of *Math Horizons* for a profile of Karen Uhlenbeck.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, T(13–16), S, L**.** *Numbers and Shapes Revisited: More Problems for Young Mathematicians.* Judita Cofman. Clarendon Pr, 1995, xi + 308 pp, $24.95 (P). [ISBN 0-19-853460-4] Sequel to *What to Solve?* (TR, January 1993). Ten chapters on such classical topics as Fibonacci numbers and generalizations, partitions, number patterns, inequalities, lattice paths and hypercubes, group theory, combinatorics. Each chapter starts with definitions, fundamental properties, and concepts followed by problems for solution. Over half the book is detailed and well-motivated solutions. Excellent resource for independent study or problem-solving seminars. LCL

General, P*, L. *The Pattern Book: Fractals, Art, and Nature.* Ed: Clifford A. Pickover. World Scientific, 1995, xxxvi + 427 pp, $59. [ISBN 981-02-1426-X] Interesting introduction to the world of patterns. In three parts: representing nature; mathematics and symmetry; human art. More than 150 patterns from science, art, and mathematics, many with computational hints and recipes for their creation. Includes glossary of technical terms. DH

Recreational Mathematics, S(13), L. *Le Sabre d'Aladin.* Gilles Cohen, Michel Criton. Federation Francaise des Jeux Mathematiques (1 Avenue Foch, 94700 Maisons-Alfort, France), 1995, 127 pp, (P). [ISBN 2-909737-08-X] 45 challenging puzzles from the Eighth International Championship of Mathematical Games. Questions are translated into English; complete solutions in French. LCL

Recreational Mathematics, S(13), L. *Le Singe et la Calculatrice.* Gilles Cohen, Michel Criton. Federation Francaise des Jeux Mathematiques (1 Avenue Foch, 94700 Maisons-Alfort, France), 1995, 128 pp, (P). [ISBN 2-909737-07-1] 47 easier puzzles from the Eighth International Championship of Mathematical Games. Questions are translated into English; complete solutions in French. LCL

Elementary, S(13). *Algebra Unplugged.* Kenn Amdahl, Jim Loats. Clearwater Pub (POB 778, Broomfield, CO 80038–0778), 1995, 258 pp, $14.95 (P). [ISBN 0-9627815-7-6] An excellent and enjoyable book. Worth having several copies around to loan to students. Some of the most noteworthy students for whom this book would be a benefit may also not be willing or able to read it. PF

Precalculus, S(12–13). *Experiments in Mathematics Using Maple.* C.T.J. Dodson, E.A. Gonzalez. Springer-Verlag, 1995, xix + 465 pp, $39 (P). [ISBN 3-540-59284-9] Paper copy of *Maple* worksheets developed by the authors for pre-calculus courses. Not a textbook. JO

Finite Mathematics, T(13–14: 1). *Finite Mathematics, Fourth Edition.* Daniel P. Maki, Maynard Thompson. McGraw-Hill, 1996, xvi + 652 pp. [ISBN 0-07-039763-5] New to this edition: additional problems; chapter on logic; additional attention to modeling techniques and problem-solving strategies. (1978 text, TR, August–September 1978; Extended Review, May 1981.) LB

Finite Mathematics, S*(13–14). *Schaum's Outline of Theory and Problems of Finite Mathematics, Second Edition.* Seymour Lip-

schutz, John J. Schiller. McGraw-Hill, 1995, vii + 483 pp, $14.95 (P). [ISBN 0-07-038002-3] Update of 1966 edition. New topics: Hamiltonian and Eulerian graphs, descriptive and inferential statistics, interest calculations. An appendix illustrates use of graphing/programmable calculator to solve sample problems. LB

Education, P*. *Schemas in Problem Solving.* Sandra P. Marshall. Cambridge Univ Pr, 1995, xi + 424 pp, $49.95. [ISBN 0-521-43072-0] Broadly, a schema is "the means by which similar experiences are assimilated and aggregated in such a way as to be quickly and easily remembered." Presents a precise operational definition, formulates a theory of schema development combining symbolic processing and connectionist learning, and illustrates five schemas by means of arithmetic story problems showing their significance for learning, instruction, and assessment. RSK

Education, P. *Teaching the Majority: Breaking the Gender Barrier in Science, Mathematics, and Engineering.* Ed: Sue V. Rosser. Teachers College Pr, 1995, vii + 264 pp, $22.95 (P); $50. [ISBN 0-8077-6276-8; 0-8077-6277-6] A series of essays providing feminist perspectives on teaching physics, chemistry, mathematics, computer science, and geoscience. The papers skewer typical approaches to inclusiveness such as Feynman's "physics is phun" or transsexual algebra problems that replace Robert with Roberta. Feminist algebra, as reported here, eschews endemic metaphors of violence (e.g., "attack the problem," "use brute force") and war (e.g., half-life of plutonium) in order to transform the "tacit misogny of the accepted mathematical paradigm." LAS

Combinatorics, S(15). *Schaum's Outline of Theory and Problems of Combinatorics.* V.K. Balakrishnan. McGraw-Hill, 1995, viii + 200 pp, $13.95 (P). [ISBN 0-07-003575-X] 800 solved problems on general concepts, generating functions and recurrence relations, and group theory. DH

Combinatorics, T(18), P, L. *The Foundations of Topological Graph Theory.* C. Paul Bonnington, Charles H.C. Little. Springer-Verlag, 1995, ix + 178 pp, $49.95. [ISBN 0-387-94557-1] Completely combinatorial approach to topological graph theory, based on 3-graph generalization of imbedding. Includes classification of surfaces, Jordan curve theorem, planar graph characterization, and duality. JH

Number Theory, T(17: 1, 2), P, L. *Introduction to Analytic and Probabilistic Number Theory.* Gérald Tenenbaum. Stud. in Adv. Math., V. 46. Cambridge Univ Pr, 1995, xv + 448 pp, $64.95. [ISBN 0-521-41261-7] A slightly expanded version of the author's 1990 French text. Includes elementary methods, the methods of complex analysis, and probabilistic methods. Excellent exercises. CEC

Number Theory, T*(14: 1, 2), S*, L.** *The Theory of Numbers: A Text and Source Book of Problems.* Andrew Adler, John E. Coury. Jones & Bartlett, 1995, xiv + 401 pp. [ISBN 0-86720-472-9] Standard introduction plus polynomial congruences, factorization of large numbers, Gaussian integers. Detailed solutions to hundreds of problems. Excellent historical references. Terse writing style. CEC

Number Theory, P, L. *An Introduction to the Smarandache Function.* Charles Ashbacher. Decisionmark (200 2nd Ave. SE, Cedar Rapids, IA 52401), 1995, 60 pp, $7.95 (P). [ISBN 1-879585-49-9] The Smarandache function $S(m)$ is the smallest value of n for which m divides $n!$. A terse, yet thorough, introduction to the elementary properties of $S(m)$. CEC

Linear Algebra, T(15–16), L. *Linear Algebra Done Right.* Sheldon Axler. Undergrad. Texts in Math. Springer-Verlag, 1996, xvii + 238 pp, $49. [ISBN 0-387-94595-4] For a second course; emphasizes theory over computation. Treats the structure of linear operators on inner product, complex, and real vector spaces without the use of determinants (which appear only in the final chapter). The determinant-free proofs are elegant and intuitive. SA

Linear Algebra, T(15–16), S, L. *Foundations of Linear Algebra.* Jonathan S. Golan. Texts in Math. Sci., V. 11. Kluwer Academic, 1995, viii + 236 pp, $94. [ISBN 0-7923-3614-3] Axiomatic approach. Develops classical linear algebra through the Jordan canonical form and the spectral theorem for complex inner product spaces. Also includes brief treatments of Moore–Penrose pseudoinverses, bilinear forms, algebras over a field. Many exercises, varying in difficulty. JS

Group Theory, T(17: 1), S, P. *The Classification of the Finite Simple Groups, Number 2.* Daniel Gorenstein, Richard Lyons, Ronald Solomon. *Part I, Chapter G: General Group Theory,* Math. Surv. & Mono., V. 40, No. 2. AMS, 1996, xi + 218 pp, $59. [ISBN 0-8218-0390-5] Continuation of the revision and simplification of the classification theorem. Covers theory needed, including semisimple subgroups, nilpotent groups, components, layers, the Fitting subgroup, signalizers, parabolic subgroups, some character theory. RM

Algebra, T(17: 1), S, P. *A Primer of Algebraic D-modules.* S.C. Coutinho. London Math. Soc. Stud. Texts, V. 33. Cambridge Univ Pr, 1995, xii + 207 pp, $22.95 (P); $54.95. [ISBN 0-521-55908-1; 0-521-55119-6] Nice, relatively self-contained introduction to the algebraic theory of modules over rings of differential operators (the Weyl algebra generated by operators, left multiplication by X_i and $\partial f/\partial X_i$ being the prototypical example). Discussion of invariants, operations, and applications (e.g., to the Jacobian conjecture, stability of differential equations). RM

Calculus, T(13: 2). *Applied Calculus with Technology.* Edmond C. Tomastik. Saunders College, 1996, xxiv + 726 pp, $31 (P). [ISBN 0-03-006863-0] Similar to *Brief Calculus* (TR, April 1996), but written for students with just two years of high school algebra. Enough material for two semesters. Incorporates graphing calculators nicely. PF

Calculus, T(13–14). *Calculus with Analytic Geometry, Second Edition.* George F. Simmons. McGraw-Hill, 1996, xxiii + 887 pp. [ISBN 0-07-057642-4] From the author: "This book is intended to be a mainstream calculus text that is suitable for every kind of course at every level. It is designed particularly for the standard course of three semesters for students of science, engineering, or mathematics." (*First Edition*, TR, December 1985.) PF

Calculus, T(13: 1). *Applied Calculus: Interpretations in Business, Life, and Social Sciences.* Denny Burzynski, Guy D. Sanders. PWS, 1996, xvii + 606 pp. [ISBN 0-534-175988] Bold effort to incorporate technology and careful interpretation of results. Worth careful consideration. PF

Real Analysis, T(16–17), S. *Lebesgue Integration, Second Edition.* Soo Bong Chae. Universitext. Springer-Verlag, 1995, xiii + 264 pp, $39.95 (P). [ISBN 0-387-94357-9] Reviews the topology of $\mathbb{R}$ and the Riemann integral, but these are really prerequisites. Develops the Lebesgue integral on a closed interval; the theory of measure follows as an application of the theory of integration. Includes a treatment of L^p spaces, with special emphasis on the L^2 theory of Fourier series. (1980 Marcel Dekker edition, TR, March 1981.) SA

Differential Equations, T*(15–16). *Introduction to Perturbation Methods.* Mark H. Holmes. Texts in Appl. Math., V. 20. Springer-Verlag, 1995, xiii + 337 pp, $39. [ISBN 0-387-94203-3] Explores the interplay between perturbation methods and scientific computing. Uses examples (ODE's, PDE's, difference equations, and transcendental equations) as motivation for technique. Topics include: multiple time scales, WKB method for linear singular perturbations, disparate spatial scales, and multiple solutions and stability. Accessible to undergraduates with a background in differential equations and elementary real analysis. MPR

Dynamical Systems, S*(17–18), P. *Introduction to the Modern Theory of Dynamical Systems.* Anatole Katok, Boris Hasselblatt. Ency. of Math. & Its Applic., V. 54. Cambridge Univ Pr, 1995, xviii + 802 pp, $79.95. [ISBN 0-521-34187-6] A comprehensive exposition. Part one focuses on fundamental examples and principle notions. Part two is dedicated to the interplay between local analysis and global complexity. Parts three and four investigate low-dimensional and hyperbolic dynamical systems. Seemingly every topic is covered in depth. Worth a look. MPR

Numerical Analysis, T(15–16: 1), S, L. *Numerical Methods.* Wolfgang Boehm, Hartmut Prautzsch. AK Peters, 1993, x + 186 pp, $39.95 (P). [ISBN 1-56881-020-2] Important numerical algorithms for matrices, optimization, iterative methods, discrete approximation and interpolation, numerical differentiation and integration. No thorough theoretical development. Nice, crisp explanations and understandable descriptions. RM

Geometry, T(13–16), S*, L*. *Circles: A Mathematical View.* Dan Pedoe. MAA, 1995, xxxvii + 102 pp, $18.95 (P). [ISBN 0-88385-518-6] Revised edition of the 1957 Dover classic (TR, June–July 1980) includes a new introductory Chapter 0 which introduces readers to special vocabulary and circle concepts needed. Exercises with solutions. LCL

Elementary Statistics, S?(13–14). *The Common Sense for a First Course in Statistics.* T.P. Hutchinson. Rumsby Scientific, 1995, 29 pp, $3 (P). [ISBN 0-646-24865-0] Tiny pamphlet resembles a series of handouts for review prior to an exam. LB

Elementary Statistics, T(13: 1). *Statistics for the Social Sciences.* Victoria L. Mantzopoulos. Prentice Hall, 1995, xv + 382 pp. [ISBN 0-13-101320-3] Standard topics. Unlike many new texts, encourages hand calculations. "A Learning Aid" box in each section facilitates step-wise calculations. Thorough coverage of probability. Numerous graphics. KS

Mathematical Statistics, T*(17: 1, 2). *Linear Statistical Models.* James H. Stapleton. Ser. in Prob. & Stat. Wiley, 1995, xiii + 449 pp, $59.95. [ISBN 0-471-57150-4] Emphasizes

the geometry of vector spaces. Includes necessary linear algebra, a variety of topics in multiple regression and analysis of variance, and a chapter on categorical data analysis using loglinear models. RSK

Statistical Methods, T(16–18). *Multivariate Data Analysis with Readings, Fourth Edition.* Joseph F. Hair, Jr., *et al.* Prentice Hall, 1995, xvii + 757 pp. [ISBN 0-02-349020-9] Covers multiple regression, discriminant analysis, MANOVA, canonical correlation, FA, cluster analysis, multidimensional scaling, conjoint analysis, and SEM. 6-stage model-building approach to each topic: objectives, design issues, assumptions, model estimation, interpretation, and validation of results. Non-technical; no linear algebra. Readings from business for each topic. Appendix with commands for SAS, SPSS, BMDP, and LISREL. RS

Statistical Methods, T*(17–18), C. *Applied Multivariate Techniques.* Subhash Sharma. Wiley, 1996, xviii + 493 pp, $75.95, with disk. [ISBN 0-471-31064-6] Covers PCA, FA, CFA, cluster analysis, discriminant analysis, logistic regression, MANOVA, canonical correlation, and SEM. For each, the following are discussed: geometric view, analytic view (at college algebra level), assumptions, interpretation of results, technical issues (using linear algebra). Annotated examples use SAS and SPSS. Data disk. RS

Statistical Methods, T(16–17: 1), C, P. *Statistical Analysis of Nonnormal Data.* J.V. Deshpande, A.P. Gore, A. Shanubhogue. Wiley, 1995, viii + 240 pp, $37.95, with disk. [ISBN 0-470-22057-0] Assumes familiarity with normal-theory statistical principles. Discusses methods for discrete data, data from specific non-normal continuous distributions, as well as standard nonparametric methods. Contains methods for reliability and survival analysis. Each method followed by a worked example and an exercise. Readable. Includes DOS disk with BASIC programs. RS

Mathematical Computing, C, P. *XploRe: An Interactive Statistical Computing Environment.* W. Härdle, S. Klinke, B.A. Turlach. Stat. & Comp. Springer-Verlag, 1995, xvi + 387 pp, $44.95. [ISBN 0-387-94429-X] Introductory guide to the matrix-oriented language *XploRe.* Includes a section on teaching statistics using interactive graphics. RS

Mathematical Computing, T(15). *Classical Mechanics with Maple.* Ronald L. Greene. Springer-Verlag, 1995, ix + 173 pp, $29 (P). [ISBN 0-387-94512-1] Uses Maple V to illuminate some problems from introductory classical mechanics: collisions, harmonic oscillators, and the N-body problem. A nice introduction to basic Maple techniques, but rarely reveals the power of computer algebra systems to investigate otherwise unapproachable problems. MPR

Computer Literacy, S. *GNU EMACS Reference.* Dennis Gentry. Specialized Systems Consultants, 1995, 18 pp, $4.50 (P). Handy-dandy little foldout card containing lots of *emacs* commands. MPR

Applications (Physics), T(17–18: 1), S, P. *Introduction to Octonion and Other Non-Associative Algebras in Physics.* Susumu Okubo. Montroll Memorial Lect. Ser. in Math. Physics, V. 2. Cambridge Univ Pr, 1995, xi + 136 pp, $49.95. [ISBN 0-521-47215-6] Overview of the development, theory, and use of non-associative algebras in physics, including algebras of physical observables in quantum mechanics, angular momentum and octonians, the Yang–Baxter equations, and Yang–Mills gauge theory. RM

Applications (Physics), T*(15–16), L. *Thermodynamics and Statistical Mechanics.* Walter Greiner, Ludwig Neise, Horst Stöcker. Springer-Verlag, 1995, xii + 463 pp, $49 (P). [ISBN 0-387-94299-8] English version of well-known series of German physics texts by the first author. Emphasizes microscopic view of thermodynamics and an understanding of macroscopic phenomena. Covers statistical mechanics, quantum statistics, real gases, and phase transitions. Many, many examples and problems. MPR

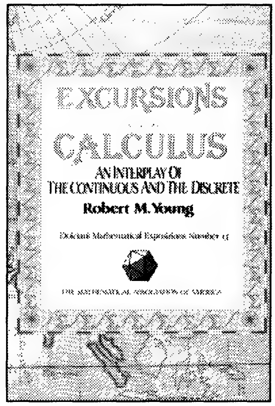
Applications (Physics), P. *Renormalization Group.* Giuseppe Benfatto, Giovanni Gallovotti. Physics Notes, V. 1. Princeton Univ Pr, 1995, viii + 143 pp, $14.95 (P); $35. [ISBN 0-691-04446-5; 0-691-04447-3] Uses the renormalization group point of view to show how physically plausible conjectures can be transformed into precise mathematical statements. Examples from constructive theory of Euclidean fields, universality theory of critical points in statistical mechanics, and the theory of Fermi surfaces in Fermi liquids. MPR

Reviewers

SA: Steve Abbott, St. Olaf; LB: Lynne Baur, Carleton; CEC: Clifton E. Corzatt, St. Olaf; PF: Paul Froeschl, Macalester; DH: Deanna Haunsperger, Carleton; JH: Joan Hutchinson, Macalester; RSK: Richard S. Kleber, St. Olaf; LCL: Loren C. Larson, St. Olaf; RM: Richard Molnar, Macalester; JO: Jeff Ondich, Carleton; MPR: Matthew P. Richey, St. Olaf; JS: John Schue, Macalester; RS: Richard Single, St. Olaf; KS: Karla Spriestersbach, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf.

EXCURSIONS IN CALCULUS: an Interplay of the Continuous and the Discrete

Robert M. Young



An excellent source of projects for well motivated students. This list of 463 references is a valuable aid for those who wish to dig deeper. *—CHOICE*

This excellent book belongs in every school library and on every mathematics teacher's bookshelf.
—The Mathematics Teacher

Any book published in the Dolciani Mathematical Expositions series will naturally be measured against the remarkably high standards already well established by the previously published titles. In the present case, however, it is safe to say that this latest addition (the thirteenth in the series) not only easily meets the previously set standards but sets an entirely new standard for future volumes.
—Mathematical Reviews

Printed with eight full–color plates.

The purpose of this book is to explore, within the context of elementary calculus, the rich and elegant interplay that exists between the two main currents of mathematics, the continuous and the discrete. Such fundamental notions in discrete mathematics as induction, recursion, combinatorics, number theory, discrete probability, and the algorithmic point of view as a unifying principle are continually explored as they interact with traditional calculus. The interaction enriches both.

The book is addressed primarily to well-trained calculus students and their teachers, but it can serve as a supplement in a traditional calculus course for anyone who wants to see more.

CONTENTS:

- Infinite Ascent, Infinite Descent: The Principle of Mathematical Induction
- Patterns, Polynomials, and Primes: Three Applications of the Binomial Theorem
- Fibonacci Numbers: Function and Form
- On the Average
- Approximation: from Pi to the Prime Number Theorem
- Infinite Sums: A Potpourri

The problems, taken for the most part from probability, analysis and number theory, are an integral part of the text. Many point the reader toward further excursions. There are over 400 problems presented in this book.

408 pp., 1992, Paperbound

ISBN 0-88385-317-5

List: $47.95 MAA Member: $39.00

Catalog Number DOL-13

ORDER FROM:

The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036
1-800-331-1622 Fax (202) 265-2384

Membership Code ______

Name ____________________

Address ____________________

City ____________________

State___ Zip Code ____________

Qty. Catalog Number Price

____

____

Total $________

Payment ❒ Check ❒ VISA ❒ MASTERCARD

Credit Card No. ____________________

Signature ____________ Exp. Date ________

She Does Math!

Real-Life Problems From Women On The Job

Marla Parker, Editor

She Does Math! presents the career histories of 38 professional women and math problems related to their work. Each history describes how much math the author took in high school and college; how she chose her field of study; and how she ended up in her current job. Each of the women presents problems that are typical of those she has faced in her job. The problems require only high school mathematics for their solution.

Who should have this book?

Your daughter, your granddaughter, your sister, your former math teacher, your students—and young men too. They want to know how the math they study is applied, and this book will show them.

There are many reasons to buy this book:

➤ By reading the career histories of the women profiled in this book, young people will learn that if they take mathematics courses in high school and college they will be qualified to enter interesting technical fields and earn good salaries.
➤ The problems have special appeal to students who are beginning to think about career choices.
➤ The book provides practical information about the job market in an interesting, innovative way.
➤ Strong female role models who work as successful technical professionals are presented.
➤ The problems are interesting and challenging, yet require only high school mathematics. They demonstrate how good math skills are applied to real-life problems.

Read what others have said about She Does Math!

Finally — a practical, innovative, well-written book that will also inspire its readers. The wonder is...it's a mathematics text and a biography! The idea of women telling their own career stories, emphasizing the mathematics they use in their jobs is extremely creative. This book makes me wish that I could go through school all over again!

Anne L. Bryant, Executive Director
American Association of University Women

She Does Math! *will undoubtedly appeal both to students who already enjoy math and want to get a view of potential career paths, and also to students who want to better understand the relevance of their math classes to their future careers. It is an absorbing look into the lives of some very inspiring and talented women!*

Susanne Hupfer and Elisabeth Freeman
Yale University

This collection is a wonderful confirmation that real women do math. They do math in a surprising variety of careers, fully enjoying the challenge and rewards of solving complex problems. This is a book for young women and men, a book for their teachers and parents, a book that informs about the possibilities that mathematics affords to all. It is also a book that will engage you in real-life mathematics! – Doris Schattschneider
Moravian College

272 pp., Paperbound, 1995 ISBN 0-88385-702-2
List: $27.50 MAA Member: $20.95
Catalog Code: SDM/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
_____	SDM/JR	_____	_____
_____	_____	_____	_____
		TOTAL	_____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

Five Hundred Mathematical Challenges

Edward J. Barbeau, William O. Moser, and Murray S. Klamkin

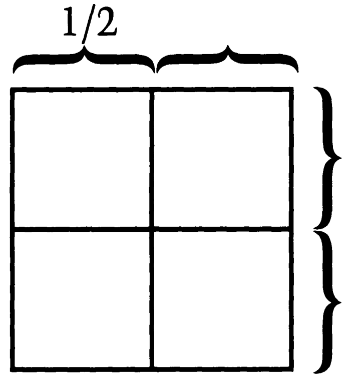
This book contains 500 problems that range over a wide spectrum of areas of high school mathematics and levels of difficulty. Some are simple mathematical puzzlers while others are serious problems at the Olympiad level. Students of all levels of interest and ability will be entertained and taught by the book. For many problems, more than one solution is supplied so that students can see how different approaches can be taken to a problem and compare the elegance and efficiency of different tools that might be applied.

Teachers at both the college and secondary levels will find the book useful, both for encouraging their students and for their own pleasure. Some of the problems can be used to provide a little spice in the regular curriculum by demonstrating the power of very basic techniques.

These problems were first published as a series of problem booklets almost twenty years ago, at a time when there were few resources of this type available for the English reader. They have stood the test of time and the demand for them has been steady. Their publication in book form is long overdue.

This collection provides a solid base for students who wish to enter competitions at the Olympiad level. They can begin with easy problems and progress to more demanding ones. A special mathematical tool chest summarizes the results and techniques needed by competition-level students.

Problem 14.
Show that if 5 points are all in, or on, a square of side 1, then some pair of them will be no further than $\frac{\sqrt{2}}{2}$ apart.



Solution.
Divide the square into four squares of side 1/2. By the Pigeonhole Principle, one of these four squares contains at least two of the points, whose distance apart must be no greater than the diagonal of the square of side 1/2, namely $\frac{\sqrt{2}}{2}$

300 pp., Paperbound, 1995
ISBN 0-88385-519-4
List: $29.50
MAA Member: $24.95
Catalog Code: CHMP/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ___ ___ ___ ___ ___ ___

Name ____________________

Address ____________________

City ____________________

State __________ Zip __________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CHMP/JR	______	______
______	______	______	______
		TOTAL	______

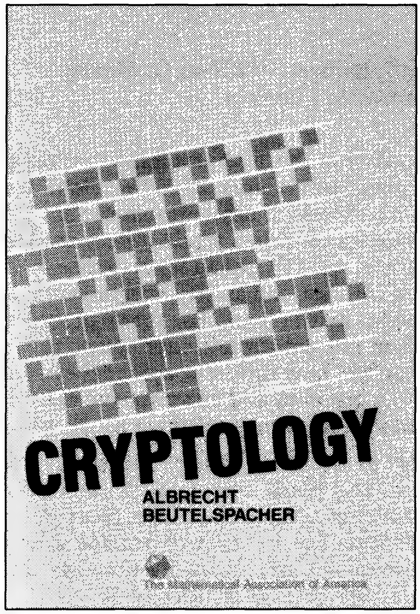
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires __/__

Signature ____________________

CRYPTOLOGY

Albrecht Beutelspacher



This fascinating little book is eminently readable, and it is a great deal of fun to peruse... the book is a real treat. We need more books like this, crafted by expert hands yet crafted so that the general reader can enjoy them.
—Bulletin of The Institute of Combinatorics and Its Applications

This excellent and entertaining book is suitable for a first course in cryptology for mathematical enthusiasts. An abundance of exercises and an excellent list of related references are included.
—The Mathematics Teacher

In spite of the light-hearted style in which the book is written throughout, it is a serious—-and successful—- attempt to explain the basis of coding and decoding messages...I can strongly recommend this book to anyone who wants a brief but comprehensive, eminently readable, and up-to-date introduction to this increasingly popular topic.
— The Mathematical Gazette

All of cryptology is covered in this work...Occupying a niche in the halls of the ivory tower of pure mathematics for nearly two millennia, number theory now forms a pillar of modern society. This book is the best explanation available today of how that pillar was constructed.
— Charles Aschbacher

A model to follow in order to make mathematics better known and understood. Accessible to a broad audience. Have fun reading this book, while you are getting a better understanding of cryptology.
— Bulletin of the Belgian Mathematics Society

How can messages be transmitted secretly? How can one guarantee that the message arrives safely in the right hands exactly as it was transmitted? Cryptology—the art and science of "secret writing"—provides ideal methods to solve these problems of data security.

The book is fun to read, and the author presents the material clearly and simply. Many exercises and references accompany each chapter.

176 pp., Paperbound, 1994
ISBN 0-88385-504-6
List: $34.00
MAA Member: $26.95
Catalog Code: CRYPT/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CRYPT/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________________ Expires ___/___

Signature ______________________

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 103, Number 6 June - July 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generally of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428B
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

All the Way with Gauss-Bonnet and the Sociology of Mathematics

Daniel Henry Gottlieb

I was stimulated to write this story by the discussion in The American Mathematical Monthly between Peter Hilton and Jean Pederson on the one hand and Branko Grünbaum and G. C. Shephard on the other hand [HP] [GS]. The discussion as well as my story involves the Euler-Poincaré Number, alias the Euler Characteristic. The *discussion* centers on whether the Euler-Poincaré Number should be discussed in a historical way without mentioning the vast and dramatic generalization and depth of understanding that this most interesting invariant has acquired in this century.

My position in this discussion is that Topology should not be viewed as an advanced subject whose theorems and concepts should be avoided until graduate school. Rather it is the study of continuity, and thus underlies the most basic geometric results. In this paper I show how the basic concept of angle leads naturally to the basic topological ideas of *degree of mapping* and of the Euler-Poincaré Number.

My *story* spans the history of mathematics. It concerns what may be the most widely known non-obvious theorem of mathematics and it contains the same stunning generalization that characterizes the recent history of the Euler-Poincaré number. In fact, it concerns one of the most important and earliest of the applications of the Euler-Poincaré number. It shows the fickleness of mathematical fame, it shows the unreasonable power of unreasonable points of view, and it shows how easy it is for mathematicians to miss and forget beautiful and important theorems as well as simple and revealing points of view.

This is a history of the Gauss-Bonnet theorem as I see it. I am not a mathematical historian. I quote only secondary sources or first hand papers that I quickly scanned, and I did not conduct any thorough interviews. Nonetheless, I am writing this history because I have contributed the last sentence to it (for the moment).

I especially want to acknowledge the help of Hans Samelson. His scholarship greatly altered the thrust of earlier versions of this paper. He discovered Satz VI. He informed me of many points in this history; about Gauss' work, Descartes work, and Hopf's work. And he was a student of Hopf who generalized the Gauss-Bonnet theorem himself.

THE NORMAL MAP. What is the most widely known, not immediately obvious, mathematical theorem? I contend that is the following: *The sum of the interior angles of a triangle equals* π. The ordinary person might admit lightly that he doesn't quite remember the Pythagorean theorem, but if he does not know the sum of the angles equals 180 degrees, he brands himself as uneducated. I will call this theorem the *180 degree theorem.*

This 180 degree theorem was proved in the time of Thales. It has undergone a remarkable generalization through the ages, culminating in the Gauss-Bonnet Theorem as I give it here. The first generalization involves the concept of exterior angle. Exterior angles contain the same mathematical information as interior angles because (see Figure 1) they are related by a simple equation: $\alpha + \beta = \pi$, where α is an interior angle and β is the corresponding exterior angle. Now *the sum of the exterior angles of a polygon equals* 2π. This immediately implies the 180 degree theorem by the preceding equation.

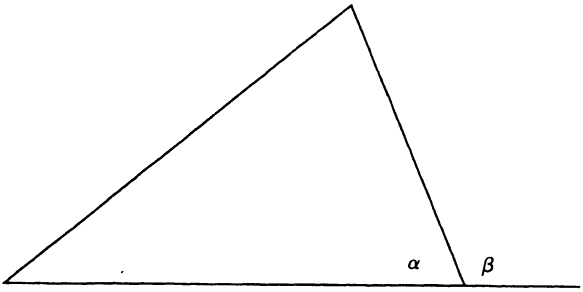


Figure 1

What is the angle between two straight lines intersecting at a point O? Let S^1 be the unit circle centered at O. Then the length of the arc of S^1 cut off by the lines (see Figure 2) is the angle between the lines. We regard angle as a property of a subset of the unit circle rather than as a number. This point of view is closer to the original Greek point of view. Regarding angle as a number is a more modern point of view.

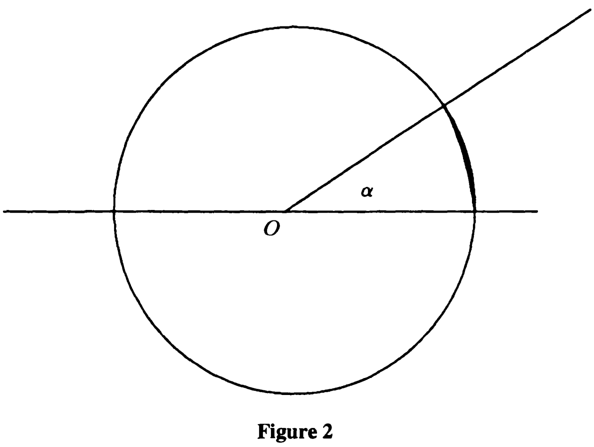


Figure 2

This Greek point of view is susceptible to immediate generalization. Just as angle is the length, or 1-volume, of a region of the unit circle in two space, we can think of the area, or 2-volume, of a region on the unit sphere in three space, denoted by S^2, as a representation of angle in three space. In general, angle in n-space can be thought of as the $(n-1)$-volume of a region on the unit sphere S^{n-1} in n-space.

Now consider a plane curve σ connecting point A to point B (Figure 3). Consider the unit vectors tangent to σ at A and B. Translate these vectors to the origin, keeping the initial and translated vectors parallel. Then the arc on S^1 cut off by the two translated vectors represents the angle the curve has turned through.

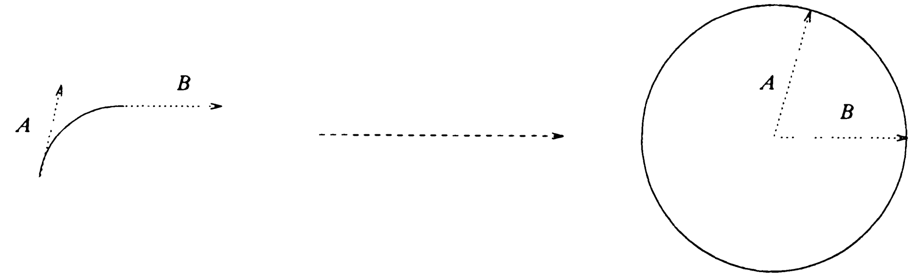


Figure 3

One thing that topologists have learned in developing Topology is that it almost always pays to convert things into functions or mappings. This procedure has spread throughout all of mathematics in the last half of this century. So in the case at hand, define a mapping from σ to the unit circle S^1 as follows: At each point P on σ, construct the unit tangent vector to σ at P, then parallel translate it to the origin; its end lies on the unit circle. Call this the *tangent map*.

Now let B approach A along σ. If we divide the angle between the tangent at B and the tangent at A by the length along σ from A to B, we have a quantity that approaches a limit if σ is smooth enough. This number is the *curvature* of σ at A. This is the same as saying that the curvature at A is the reciprocal of the ratio at A of the length of an infinitesimal arc on σ to the length of its image on S^1.

Now let us approximate a polygon by a smooth simple closed curve. Then the rate of change of the tangent (the *curvature* of the curve) corresponds to the exterior angle, and the total turning of the tangent (the *total curvature* of the closed curve) corresponds to the sum of the exterior angles. Now for simple closed curves, the tangent turns through 2π as it completes a tour of the closed non-self intersecting curve. That is, the total curvature is 2π. This then implies the exterior angles sum to 2π by continuity. This approximation of polygons by smooth curves is an argument known to the Greeks. So we have greatly generalized the original 180 degree theorem about the triangle by the theorem that the total curvature of a simple closed curve is 2π.

Instead of the tangents, we could consider the normals to σ. The normal varies exactly as the tangent does as a point moves along σ, so we could define the curvature of σ using normals instead of tangents. Thus we replace the tangent map with the *normal map* from σ to S^1. The advantage of using normals instead of tangents is that we can generalize curvature to surfaces in three space, for on surfaces in three-space, the normal direction is well-defined whereas there is no unique tangent direction.

We formalize this concept by introducing the idea of the *Gauss map*, also called the *normal map*. To each point on a smooth surface in three-space one can assign a unique unit normal vector pointing outside. This mapping maps the surface to the unit sphere. It is given by sending each point to its normal vector and then parallel transporting the unit vector through space so that the beginning of the

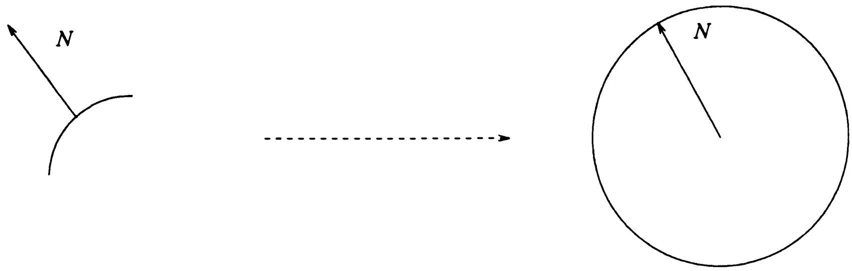


Figure 4

vector is at the center of the unit sphere and then taking the point on the unit sphere that corresponds to the tip of the transported unit vector (see Figure 4).

The same idea gives the normal map in dimension two from a closed curve to the unit circle, and from a smooth closed $(n-1)$ dimensional manifold M embedded in n-dimensional Euclidean space R^n to the unit sphere S^{n-1}. We let $\gamma : M \to S^{n-1}$ denote the normal map.

CURVATURE. Now we can define the concept of *normal curvature* at a point m of M in R^n. Let R be a small region around m in M. Let $\gamma(R)$ denote the image in S^{n-1} of R. Then the *normal curvature* at m, denoted $K(m)$, is the limit as R tends to m of the $(n-1)$ volume of $\gamma(R)$ divided by the $(n-1)$ volume of R. This is given a positive sign if γ preserves the orientation at m and a negative sign if γ reverses the orientation at m. In a suitable coordinate system, $K(m)$ is the Jacobian of γ at m.

Just as the reciprocal ratio of infinitesimal length at x on a curve to the length at the image $\gamma(x)$ is the definition of curvature of a curve in the plane at x, so is the reciprocal ratio of infinitesimal areas from x to $\gamma(x)$ the curvature of a surface at x in space. One would think that the same name would hold for the higher dimensional examples of ratios of infinitesimal volumes, but for historical reasons this did not happen. For the purposes of this paper I will call this number the *normal curvature* of M at x in R^n.

Let us pause and consider the reason that normal curvature, the natural generalization of angle, is not called curvature in dimensions higher than 2. It is because in dimension two, the normal curvature depends not on how the surface sits in R^3, but on the intrinsic geometry of the surface. That is, the curvature can be calculated by considering only the surface and not the ambient space. This is the famous Theorema Egregium of Gauss. So for higher dimension, curvature means the Riemann curvature tensor. This is based on the two dimensional curvature and does not agree at all with the normal curvature in higher dimensions and does not even make sense for dimension 1 curves. This curvature tensor plays an important role in differential geometry and physics, but it does not replace the normal curvature the way interior angles are replaced by exterior angles. Outside of dimension 2 they are very different concepts. This issue of intrinsic vs. extrinsic will play a key role in my story.

Now consider a compact $(n-1)$-dimensional manifold M in R^n, and assume that M has no boundary. Now M divides R^n into two pieces, the interior and the exterior. Let N denote the interior of M, which is a manifold with boundary M. Now if we integrate the normal curvature K over M, we get $\int K dM$, the analogue

of the sum of the exterior angles. Call this the *Total Curvature* or the old fashioned *Curvatura Integra* of M in R^n. Now we can state our version of the Gauss-Bonnet theorem. Here the Euler-Poincaré number of N is $\chi(N)$.

Gauss-Bonnet Theorem. $\int K\,dM = \chi(N) \times$ (*the volume of* S^{n-1})

NORMAL DEGREE. The unit volume of the $(n-1)$-sphere is 2π for the 1-sphere and 4π for the 2-sphere and it changes form for each dimension. Thus we define the *degree* of γ by the Curvatura Integra divided by the volume of the unit sphere corresponding to the dimension of M. The degree of γ is denoted by $\deg(\gamma)$ and is called the *normal degree*. The normal degree turns out to be an integer. In fact, this is a special case of the concept of *degree of a mapping*, an integer that plays a major role in Topology. In this notation we can write the Gauss-Bonnet theorem as the

Gauss-Bonnet-Hopf Theorem. $\deg(\gamma) = \chi(N)$.

The Euler-Poincaré number is the earliest invariant of Algebraic Topology. It is a vast generalization of a formula involving convex polyhedra due to Euler. There is evidence that Descartes knew about this formula a century before Euler, [S_2] or [St].

The degree of a map can be traced back to Kronecker and was well understood by L. E. J. Brouwer around 1913. The integral definition given here for the Gauss map can be generalized to maps between oriented closed manifolds of the same dimension. The most general definitions of the degree of a mapping and of the Euler-Poincaré number require Homology Theory. But both these concepts were discovered before homology was well understood and they can be used very effectively without knowledge of homology.

For a two dimensional surface N that can be divided up nicely by triangles that fit nicely together to form what we call a triangulation (as in Figure 5), the Euler-Poincaré number satisfies

$$\chi(N) = v - e + f$$

where v is the number of vertices, e is the number of edges, and f is the number of triangles in the triangulation.

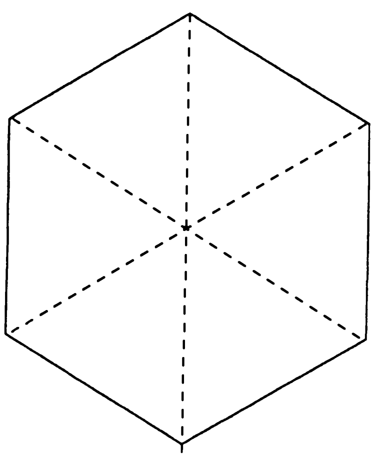


Figure 5

Given this, it is a simple matter to show that if N is bounded by a convex polygon, then $\chi(N) = 1$. Hence $\deg(\gamma) = 1$ by the Gauss-Bonnet-Hopf theorem so $\int K dM = 2\pi$, where K denotes the curvature of the curve in the plane. As we have said, this gives the 180 degree theorem.

Thus we have a tremendous generalization of the sum of angles concept valid for every dimension and given by a simple formula. We continue with the remarkable history of this result.

THE NINETEENTH CENTURY. The Gauss-Bonnet Theorem is so interesting that various authors could not resist including parts of its history in their textbooks. For example, Spivak [Sp] and Stillwell [St] give accounts of its early history.

Consider a geodesic triangle T on a surface in three space. The edges of the triangle are geodesics. Geodesics are what passes for straight lines on the surface; they are paths of shortest length on the surface. Let α, β, γ denote the interior angles of the triangle (Figure 6). Then if we integrate the curvature K over the triangle T, we get the Gauss-Bonnet Formula:

Gauss-Bonnet Formula for the geodesic triangle. $\int K dT = \alpha + \beta + \gamma - \pi$

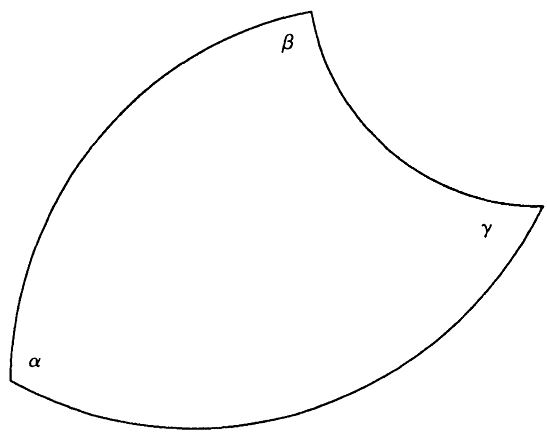


Figure 6

This formula immediately gives interesting corollaries:

If the triangle T is a plane triangle, then the geodesics are straight lines and K is identically equal to zero, so $\alpha + \beta + \gamma = \pi$. So the Gauss-Bonnet Formula implies the 180 degree theorem, but not at all in the same way that the Gauss-Bonnet-Hopf Theorem implies the 180 degree theorem.

If we divide the angular excess $\alpha + \beta + \gamma - \pi$ by the area of T, we get a number that is calculated intrinsically on the surface. As we let T shrink down to a point m, the ratio approaches the curvature $K(m)$ at m. Hence K is an intrinsic concept of the surface. This is Gauss' famous Theorema Egregium, but his published proof is not the argument just given. In an earlier unpublished manuscript, he gave this argument right after his proof of the Theorema Egregium.

If we triangulate a closed surface M with geodesic triangles, we get a Gauss-Bonnet formula for each triangle. If we add these equations up, we get on the left side the *total curvature* (also called the Curvatura Integra): $\int K dT$. On the right side we can rearrange the angles cleverly and end up with $4\pi \times \chi(M)/2$.

This agrees with what we named the Gauss-Bonnet Theorem, because for surfaces $\chi(M) = 2 \times \chi(N)$, where N is the part of space interior to the closed surface M. In fact, it is true that $\chi(M) = 2 \times \chi(N)$ for all even dimensional M. For odd dimensional closed manifolds M, however, $\chi(M) = 0$. These elementary topological facts along with 'intrinsic vs. non-intrinsic' play a key role in this story.

Gauss wrote down the preceding version of the 'Gauss-Bonnet Formula for the geodesic triangle' in an unpublished manuscript in 1825. In 1827, he published a book giving a differential formula, which if integrated would have given the generalization that Bonnet got of the Gauss-Bonnet formula; I was informed of this by Samelson.

In 1848, O. Bonnet extended the Gauss-Bonnet formula for a triangle to smooth closed curves on the surface. Here the sum of the angles is essentially replaced by the integral of the *geodesic curvature*. This generalized formula acquired the name Gauss-Bonnet sometime later. Probably Blaschke was the first to use the name in a textbook in the early 1920's.

If the geodesic triangles triangulate a closed surface S that is topologically a sphere, then Euler's Formula

$$v - e + f = 2$$

gives the first global Gauss-Bonnet theorem: $\int K dS = 4\pi$.

A lost manuscript of Descartes copied in Leibniz' hand was discovered some years later and published in the Comptes Rendus in 1860. A note by Bertrand immediately following Descartes' article points out its relationship to the global theorem. Bertrand notes that Descartes seems to get the polyhedral version of the global Gauss-Bonnet Theorem. He attributes the global theorem to Gauss. See [S_2] for an interesting account of this manuscript. However, we know that nobody understood the Euler-Poincaré Number at that time, and the result really held only for a surface diffeomorphic to a sphere. A good account of the difficulty involved with the development of the Euler-Poincaré Number is found in [La]. Indeed, the Hilton et al. discussion would fit right into the dialogues that Lakatos used to present his thesis.

Walter Dyck seems to be the first to realize that the Gauss-Bonnet Theorem should hold for more that just spherical surfaces. He did this in 1888. According to Hirsch [Hi], Dyck was the first to connect the degree with the Euler-Poincaré number and thus prove "what is wrongly called the Gauss-Bonnet Theorem".

An examination of Dyck's paper reveals pictures that are reminiscent of standard figures in Morse Theory, developed 50 years later. Dyck was a real pioneer, but he, like Descartes, was ahead of his time. Samelson tells me that he cannot find a statement of the global Gauss-Bonnet theorem in Gauss' works. So it appears that the global Gauss-Bonnet theorem should be called the Descartes-Dyck theorem.

Actually, part of this story shows that the name of a theorem is not really for an attribution. It is very convenient to have a name for important theorems, and the main point is that people should know approximately what theorem is meant by the name rather than who gets the credit. Still, one can reflect that Bonnet's name is famous and Dyck's is virtually unknown these days.

HOPF TO CHERN. Dyck worked at a time when two basic ideas—degree of a map and the Euler-Poincaré number—were not clearly understood. By 1925, these

concepts were well-defined and were found to be useful. This was due in no small measure to Heinz Hopf.

Hopf made the biggest advance in [H_1]. He essentially proved that $\deg(\gamma) = \chi(M)/2$ for closed hypersurfaces of *even* dimension. The factor $1/2$ is explained by the fact that $\chi(N) = \chi(M)/2$ whenever N is a compact odd-dimensional manifold with boundary M. Since $\chi(M) = 0$ for closed odd-dimensional manifolds, the theorem as stated by Hopf did not seem to generalize to the odd dimensional case, and in particular did not generalize the 180 degree theorem, which as we saw *is* generalized by the Gauss-Bonnet Formula.

Since the curvature of a surface is intrinsic in dimension 2, Hopf asked for intrinsic proofs and generalizations of his result [H_3]. He did this repeatedly and interested several mathematicians in the question. The story is told in [Gr].

Using Hermann Weyl's theory of tubes, two mathematicians independently answered Hopf's question in 1940. Allendoerfer [Al] and Fenchel [Fe] discovered that $\deg(\gamma)$ of the boundary of a tubular neighborhood of a closed $2n$ dimensional manifold embedded in a $2r$ dimensional Euclidean Space is equal to the integral of a $2n$ form constructed out of the components of the Riemannian curvature tensor and combined together as a Pfaffian. All this is too complicated to describe here. Since the tubular neighborhood has the same Euler-Poincaré Number as the embedded manifold, they got a formula for the Euler-Poincaré Number in terms of the Riemannian curvature of an embedded even dimensional manifold. This remarkable formula held for every Riemannian manifold because every Riemannian manifold can be isometrically embedded in some Euclidean space. However, this last result was not known until the 1950's, when it was proved by Nash.

Although the Allendoerfer-Fenchel Formula held only for an embedded manifold, it was obviously independent of the embedding and begged for an intrinsic proof. S. S. Chern provided one in 1944 [Ch]. This proof was so well received that the Allendoerfer-Fenchel Formula is frequently called the Gauss-Bonnet-Chern Formula or the Gauss-Bonnet-Chern Theorem. In fact, one of the goals of Gray's book [Gr] was to prevent the interesting methods of the Tube proof from being totally submerged by the powerful ideas of Chern's proof.

SATZ VI. Now we come to the most interesting part of the story. In 1956, Hopf gave lectures on global differential geometry at Stanford University. These lectures were honored by being published as volume number 1000 of Springer-Verlag's Lecture Notes In Mathematics in 1983 [H_4]. On pages 117–118, Hopf describes his version of the Gauss-Bonnet theorem for even dimensions. He does not mention the part that holds for odd dimensions. Because of this and various conversations, I wrote the following three paragraphs.

> It is clear that at that time Hopf did not know that the Gauss-Bonnet theorem held for all dimensions and thus was a generalization of the 180 degree theorem. Or else he knew it, but was embarassed to state it. Hopf certainly knew all the ingredients for the proof in all dimensions for many years, and the proof is of the same order of difficulty as his even dimensional proof. Had he known the version that held for all dimensions it seems likely he would not have asked for intrinsic proofs, since there are none in odd dimensions. So two very fruitful lines of research probably would not have been undertaken.
>
> Yet the Gauss-Bonnet-Hopf theorem was known to several topologists around the mid fifties, among them Milnor and Lashof. Nobody seems to

know who it was who first stated the theorem. At the time there were sophisticated generalizations and studies of deg(γ), for example [Ke] and [Mi]. Just recently Bredon, in his textbook [Br], stated and proved the result as "Theorem 12.11 (Lefschetz)". He proves it as a corollary of the Lefschetz fixed point theorem.

Finally, in 1960, the Gauss-Bonnet-Hopf theorem was stated in the literature, but in an even more generalized form by Samelson [S_1] and Haefliger [Ha]: Let N be a compact n-dimensional manifold with boundary M and let $f: N \to R^n$ be an immersion. Then the Gauss map $\gamma: M \to S^{n-1}$ can still be defined and $\deg(\gamma) = \chi(N)$.

After those words were written I received a letter from Hans Samelson. I had asked Samelson if he knew who had first discovered the Gauss-Bonnet-Hopf Theorem. After all, he had generalized it in [S_1]. In addition, he is a scholar about the Gauss-Bonnet Theorem, and he was a student of Heinz Hopf!

He thought it was Morse who first stated it. He could not find the reference, but on a hunch he looked at Hopf's 1927 paper [H_2]. There on page 248, Satz VI, the Gauss-Bonnet-Hopf theorem is clearly stated for all dimensions!

It is a testament to Hopf's genius that even though he knew Satz VI for all dimensions, the fact that the even dimensional case was true for immersions, instead of merely embeddings (concepts not well understood then), must have led him to conjecture that there was an intrinsic proof in the even dimensional case.

A differentiable map between two manifolds of the same dimension is an *immersion* if the Jacobian of the map is not zero anywhere. It is an *embedding* if in addition the map is one-to-one. Thus, immersions are one-to-one in small neighborhoods of any point, whereas embeddings are globally one-to-one. This distinction generalizes to any mappings.

Now Satz VI, that is, the Gauss-Bonnet-Hopf theorem, was proved only for embeddings, whereas Hopf knew from [H_1], that for M an even dimensional manifold, $\deg(\gamma) = \chi(M)/2$ was true if M is immersed in Euclidean space of codimension one, i.e., the dimension of the Euclidean space is one higher than the dimension of M. By the way, since locally M is embedded in Euclidean space, there is a normal direction and so the Gauss map γ is still defined.

The distinction between the odd and even dimensional cases can be explained to anyone. A circle can be immersed in a plane with arbitrary normal degree, but a

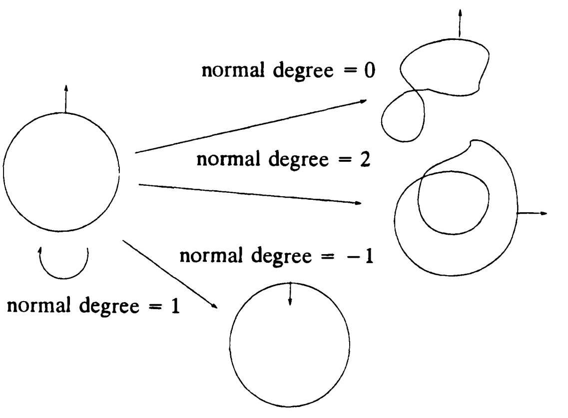


Figure 7

two-sphere can be immersed in three-space only with normal degree equal to one.

So Hopf's proof in [H_1] was not rendered superfluous by his proof of Satz VI, and he recognized the presence of intrinsicness in the difference. Thus he stimulated the Geometers with [H_3] to seek an intrinsic proof of the even dimensional Gauss-Bonnet-Hopf theorem. He also asked questions about the possible normal degrees of immersion for odd dimensional M. This stimulated Milnor's beautiful paper [Mi], and then [BK], wherein it is shown that the normal degree can take on the value of any odd integer.

THE UNASKED FOR ANSWER. In hindsight, we see that Hopf's question amounted to: Find a formula giving $\chi(M)$ in terms of the curvature tensor for even dimensional closed Riemannian manifolds. The more reasonable question should have been: Find a formula giving $\deg(\gamma)$ for all dimensions. Nobody asked this question. An answer has been found, however. It is what I will call the Topological Gauss-Bonnet Theorem to distinguish it from the Gauss-Bonnet-Chern Theorem.

This theorem immediately gives a proof of Satz VI as well as a proof for the immersion portion of the even dimensional part proved in [H_1]. The proof of this theorem requires nothing that was unknown in 1929. It is completely extrinsic. If Hopf had discovered this proof, it is unlikely he would have asked for an intrinsic proof of [H_1], and so some very important mathematics would not have been discovered so quickly. For the Allendoerfer-Fenchel Formula, now known as the Gauss-Bonnet-Chern Theorem, could not have been discovered by accident. It is too complicated. Very talented mathematicians were looking for it explicitly. On the other hand, the Topological Gauss-Bonnet Theorem is simple enough that it could have been discovered by accident. And it was!

Topological Gauss-Bonnet Theorem. *Let $f : N \to R^n$ be a map whose Jacobian is nonzero on the oriented boundary M of a compact n-manifold N. Then if x is the projection of R^n to some x-axis and $\nabla(x \circ f)$ is the gradient vector field of the composition of maps $(x \circ f)$ and Ind is its index, we have*

$$\deg(\gamma) = \chi(N) - \operatorname{Ind}(\nabla(x \circ f))$$

The fact that f has nonzero Jacobian on the boundary M of course means that f is an immersion on M. Since the composition $(x \circ f)$ is a map from N to the real line R, the gradient can be defined as in advanced calculus and gives a vector field on N. The *index* of a vector field, which is a new term in this paper, is another topological invariant that predates the start of Algebraic Topology. It was defined for vector fields in two dimensions by Poincaré in the late nineteenth century. Hopf generalized the index of a vector field for any dimensional manifold, and used the concept in his proofs of the Gauss-Bonnet-Hopf Theorem in [H_1] and [H_2].

The index of a vector field V is an integer. It is closely related to the degree of a map, yet it was defined earlier than that concept. In contrast to the degree of a map, the best definition of index does not necessarily need homology theory. In fact, it can be defined by means of a simple identity.

In 1929, Marston Morse [Mo] discovered a beautiful equation involving the index of a vector field V on a compact manifold N with boundary M; I call Morse's equation the Law of Vector Fields.

The Law of Vector Fields. *Let V be a vector field defined on N, and suppose V is not zero on the boundary M. Then* $\operatorname{Ind} V + \operatorname{Ind} \partial_- V = \chi(N)$, *where $\partial_- V$ is a vector field induced by V and defined on that part of the boundary M where V points inside.*

The vector field $\partial_- V$ is induced by V by considering the component vector field of V that is tangent to the boundary. Since $\partial_- V$ is defined on a one dimension lower space, part of the boundary M of N, an inductive scheme of calculating the index suggests itself. In fact, the Law of Vector fields is literally a self contained definition of the Index of vector fields by induction [gS]. This is elementary, but tricky, topology. Nonetheless, the whole theory of $\operatorname{Ind}(V)$ spins out from this simple 'A plus B equals C' equation. This equation is the key to the last part of the story.

Among the facts that follow easily from the Law of Vector Fields are two well known properties of the index, which combine with the Topological Gauss-Bonnet Theorem to give all the previous global results labled Gauss-Bonnet:

If V is a vector field with no zeros, then $\operatorname{Ind} V = 0$.

If V is a vector field on an odd dimensional manifold, then $\operatorname{Ind}(-V) = -\operatorname{Ind}(V)$ where $-V$ is the vector field in which every vector of V is reversed.

The Gauss-Bonnet-Hopf Theorem follows immediately from the first property, since if f is an embedding, the vector field $\nabla(x \circ f)$ is just ∇x, that is, a constant vector field parallel to the x-axis restricted to N. This has no zeros, so applying the Topological Gauss-Bonnet with the index zero yields the Gauss-Bonnet-Hopf Theorem. In fact, if f is an immersion, the vector field $\nabla(x \circ f)$ still has no zeros (because $x \circ f$ has no critical points). So we get Samelson and Haefliger's generalization of Gauss-Bonnet-Hopf from embeddings to immersions.

On the other hand, Hopf's first version in [H_1], that for even dimensional M *immersed* in R^{n+1} we have $\deg(\gamma) = \chi(M)/2$, follows from the second property. If we choose the x-axis in the Topological Gauss-Bonnet Theorem to run in the opposite direction, we reverse the direction of the gradient. The other two terms in the Topological Gauss-Bonnet Theorem certainly do not care which way the x-axis is going. So we must have $\operatorname{Ind}(\nabla(x \circ f)) = 0$. Thus $\deg(\gamma) = \chi(N) = \chi(M)/2$. The last equality follows because the Euler-Poincaré number for an even dimensional boundary is twice the Euler-Poincaré number of its bounded manifold.

There is one point that remains to be clarified. Does every orientable M that can be immersed in a codimension 1 Euclidean space bound an N so that the immersion can be extended to an f? The answer is yes. But I must admit that my way of proving this fact is immediate from a famous result of Thom's involving cobordism theory and Stiefel-Whitney numbers, which was not available until the 1950's.

THE ACCIDENTAL DISCOVERY. The Law of Vector Fields was discovered by Morse in 1929 [Mo]. In an interesting parallel with Satz VI, Morse rarely referred to the result or exploited its potential. Maybe it was because he was inventing Morse theory and may have thought unconsciously, as many topologist have, that all vector fields come from gradient vector fields. At any rate, this result was not used much and was virtually forgotten. When I rediscovered it in the 1980's, it took almost a year of questioning before someone told me about [Mo].

Ten years ago I shared the common misconception about how mathematics is created. I did not know the lessons of this story or of history. So I was shocked to

find that most topologists were unaware of what I regarded as an elementary relationship satisfied by two classical topological concepts: index and Euler-Poincaré number. So I thought, perhaps, there might be some interesting unknown consequences of the Law of Vector Fields.

I thought of a simple scheme to try to exploit the Law of Vector Fields. I looked at interesting vector fields and plugged them into the equation. I had some success with various choices. When I plugged in what I called pullback vector fields, which generalize gradient vector fields, I got an equation involving the normal degree and the Euler-Poincaré number [G_1] [G_2]. It took a while before it occurred to me that I had generalized the Gauss-Bonnet Theorem. A simplified version of that result is the Topological Gauss-Bonnet Theorem as stated above. The only simplification is that I stated the result here for gradients since it is a concept familiar from advanced calculus. In fact, pullback vector fields may even be easier than gradients.

CONCLUSIONS. Mandelbrot, in proposing the name "fractals", complained that mathematicians do not give names to concepts and results. He was right. In the deepest sense, this story really revolves about the naming of theorems and of curvature.

But it also demonstrates that several of the bromides we have grown up with are seriously flawed: That great men do not overlook simple points. That there are no great results found in using old methods. That you can't discover something good unless you have asked the right questions. That mathematics progresses mostly by the work of a few great mathematicians; this particular misconception is called the Matthew Effect by historians of Science. It seems to me that what happened with the Gauss-Bonnet Theorem happened very frequently with the best of our mathematical ideas. Nobody seems to know who invented "Cartesian Coordinates", or who first thought about higher dimensional spaces. Great mathematicians are quoted denigrating ideas that blossomed and dominated mathematics. From our present vantage point, these ideas seem trivial, but our greatest predessesors had trouble grasping them. What seems to be trivial now was once the most difficult part of mathematics: infinity, velocity and acceleration, arbitrary axioms, abstract groups, functions.

Finally, the story shows how mathematical challenges can have a great and good effect on the development of mathematics, even if the challenges were based on faulty points of view.

As an application of this last lesson I will issue a mathematical-historical challenge. Let us agree that a theorem generalizes a second theorem if the second has a short proof in which the first plays the predominate part. Then I propose the *Historical Fame Score* of any Theorem: The HFS is the product of three numbers H, F, and S.

H is the percent of the history of Mathematics covered between the time the first interesting special case was proved and its generalizing theorem was proved. The beginning of the History of Mathematics will be taken to be 300 BC in honor of Euclid and the unavailability of precise dates before that time.

F is the percent of mathematicians who know the most famous special case of the generalizing theorem.

S is the percent of results closely related to the subject matter of the generalizing theorem that receive new proofs, or new insights, or new corollaries from the generalizing theorem.

The maximum score is one million. I estimate that the Topological Gauss-Bonnet Theorem receives the maximum score. The challenge is to find generalizations with comparable scores.

REFERENCES

[Al] C. B. Allendoerfer, *The Euler Number of a Riemannian manifold*, Amer. J. Math. **62** (1940), 243–248.

[Br] Glenn E. Bredon, *Topology and Geometry*, Springer-Verlag: Graduate Text in Mathematics, New York (1993).

[BK] G. E. Bredon and A. Kosinski, *Vector fields on Π-manifolds*, Annals of Mathematics (2) **84** (1966), 85–90.

[Ch] S. S. Chern, *A simple intrinsic proof of the Gauss-Bonnet theorem for closed Riemannian manifolds*, Annals of Math. (2) **45** (1944), 747–752.

[Fe] W. Fenchel, *On the total curvature of Riemannian manifolds, I*, J. London Math. Soc. **15** (1940), 15–22.

[G_1] Daniel H. Gottlieb, *On the index of pullback vector fields*, Springer-Verlag, Lecture Notes in Mathematics **1350** (1987), 167–170.

[G_2] Daniel H. Gottlieb, *Zeroes of pullback vector fields and fixed point theory for bodies*, Contemporary Mathematics, Amer. Math. Soc. **96** (1989), 163–179.

[Gr] Alfred Gray, *Tubes*, Addison-Wesley, Redwood City, California (1990).

[gS] D. H. Gottlieb and G. Samaranayake, *Index of discontinuous vector fields*, New York Journal of Mathematics (to appear).

[GS] B. Grünbaum and G. C. Shephard, *A new look at Euler's Theorem for polyhedra*, Amer. Math. Monthly **101** (1994), 109–128.

[Ha] Andre Haefliger, *Quelques remarques sur les applications differentiables d'une surface dans le plan*, Ann. Inst. Fourier **10** (1960), 47–60.

[HP] P. J. Hilton and J. Pederson, *Euler's Theorem for polyhedra: A Topologist and a Geometer respond*, Amer. Math. Monthly **101** (1994), 959–962.

[Hi] Morris W. Hirsch, *Differential Topology*, Springer-Verlag, New York (1976).

[H_1] Heinz Hopf, *Über die Curvatura integra geschlossener Hyperflächen*, Math. Ann. **95** (1925), 340–376.

[H_2] Heinz Hopf, *Vektorfelder in n-dimensionalen Mannigfaltigkeiten*, Math. Annalen **96** (1927), 225–250.

[H_3] Heinz Hopf, *Differential Geometrie und topologische Gestalt*, Jahresbericht der Deutcher Math. Verein. **41** (1932), 209–229.

[H_4] Heinz Hopf, *Differential Geometry in the Large: Seminar Lectures NYU 1946 and Stanford 1956*, Lecture Notes in Mathematics, Springer Verlag **1000** (1983).

[Ke] M. E. Kervaire, *Courbure integrale generalisee et homotopie*, Math. Ann. **131** (1956), 219–252.

[La] Imre Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*, Cambridge University Press, Cambridge (1976).

[Mi] J. W. Milnor, *On the immersion of n-manifolds in $(n + 1)$-dimensional space*, Commentarii Math. Helvetici **30** (1956), 275–284.

[Mo] Marston Morse, *Singular points of vector fields under general boundary conditions*, Amer. J. Math **51** (1929), 165–178.

[S_1] Hans Samelson, *On immersion of manifolds*, Canadian J. Math. **12** (1960), 529–534.

[S_2] Hans Samelson, *Descartes and Differential Geometry*, Geometry, Topology, and Physics for Raoul Bott; ed. S. T. Yau; International Press, Boston (1995), 323–328.

[St] John Stillwell, *Mathematics and its History*, Springer-Verlag, New York (1974).

[Sp] Michael Spivak, *A Comprehensive Introduction to Differential Geometry*, Publish or Perish, 2nd ed, vol. 1–5, Houston (1979).

Department of Mathematics
Purdue University
West Lafayette, IN 47907
gottlieb@math.purdue.edu

When Does $A^*A = B^*B$ and Why Does One Want to Know?

Roger A. Horn and Ingram Olkin

1. INTRODUCTION. For complex scalars a and b, $\bar{a}a = \bar{b}b$ if and only if $b = e^{i\theta}a$ for some real θ. For complex or real matrices A and B, when is $A^*A = B^*B$?

In the following, we denote the m-by-n complex (respectively, real) matrices by $M_{m,n}$ (respectively, $\mathscr{M}_{m,n}(\mathbb{R})$) and we write $\mathscr{M}_{n,n} \equiv \mathscr{M}_n$. For two matrices A and B with the same number of rows, $[A\ B]$ denotes the block matrix obtained by concatenating the columns of A and B; similar notation is used for other types of block matrices. The conjugate transpose of $A \in \mathscr{M}_{m,n}$ is denoted by $A^* \equiv \bar{A}^T$, and the identity and zero matrices in $\mathscr{M}_n$ are denoted by I_n and 0_n, respectively. For $X \in M_{m,n}$, writing $X^*X = I_n$ is a short way to say that the columns of X are orthonormal. A given matrix $A \in \mathscr{M}_n$ is *Hermitian* if $A = A^*$; a real Hermitian matrix is *symmetric*. A given Hermitian $A \in \mathscr{M}_n$ (respectively, a symmetric $A \in \mathscr{M}_n(\mathbb{R})$) is *positive definite* if $x^*Ax > 0$ for all nonzero $x \in \mathbb{C}^n$ (respectively, all nonzero $x \in \mathbb{R}^n$), and is *positive semidefinite* if $x^*Ax \geq 0$ for all $x \in \mathbb{C}^n$ (respectively, all $x \in \mathbb{R}^n$). A matrix $A = [a_{ij}] \in \mathscr{M}_{m,n}$ is said to be *diagonal* if $a_{ij} = 0$ for all $i \neq j$, and is said to be *nonnegative* if all its entries are real and nonnegative. For a nonnegative diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathscr{M}_n$, we write $\Lambda^{\frac{1}{2}} \equiv \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ with nonnegative square roots. If $C \in \mathscr{M}_n$ is Hermitian and positive semidefinite, and if $C = U\Lambda U^*$ is a spectral decomposition for some unitary $U \in \mathscr{M}_n$ and nonnegative diagonal $\Lambda = \Lambda_r \oplus 0_{n-r}$ with $\Lambda_r = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ and $\lambda_i > 0$ for $i = 1, \ldots, r$, then $C^{\frac{1}{2}} \equiv U\Lambda^{\frac{1}{2}}U^*$ is the *unique* Hermitian positive semidefinite square root of C [2, Théorème 9°], [12, Theorem 7.2.6] and its *Moore-Penrose generalized inverse* is $C^{\dagger} \equiv U(\Lambda_r^{-1} \oplus 0_{n-r})U^*$; if C is real, U may be taken to be real orthogonal and $C^{\frac{1}{2}}$ and $C^{\dagger}$ are therefore real.

2. THE SINGULAR VALUE DECOMPOSITION. The following *singular value decomposition* is arguably the most basic and useful factorization known for all real or complex matrices. The proof we offer relies on the spectral theorem for Hermitian positive semidefinite matrices ([2, 6°, p. 145], [12, Theorem 2.5.6]) and the following facts:

(i) Any matrix of the form A^*A is positive semidefinite ([2, 4°, p. 144], [12, Theorem 7.2.7]),

(ii) The eigenvalues of a Hermitian positive semidefinite matrix are nonnegative ([2, 5°, p. 144], [12, Lemma 2.5.7]), and

(iii) An orthonormal set of vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$ can be extended to an orthonormal basis.

Theorem 2.1. *Let $A \in \mathscr{M}_{m,n}$ be given and let $q = \min\{m, n\}$.*

(*a*) *Suppose $A^*A = W\Lambda W^*$ for a given unitary $W \in \mathscr{M}_n$ and a nonnegative diagonal $\Lambda \equiv \operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2) \oplus 0_{n-r}$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Set* $\Sigma_r \equiv \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$

and write $W = [W_r\ Y]$ with $W_r \in \mathcal{M}_{n,r}$, so $W_r^* W_r = I_r$. There is a $V_r \in \mathcal{M}_{m,r}$ such that $V_r^* V_r = I_r$ and $A = V_r \Sigma_r W_r^*$. If A is real, then V_r and W_r may be chosen to be real.

(b) There exist unitary matrices $V \in \mathcal{M}_m$ and $W \in \mathcal{M}_n$ and a unique nonnegative diagonal matrix $\Sigma(A) = [\sigma_{ij}] \in \mathcal{M}_{m,n}$ such that $\sigma_{11} \geq \cdots \geq \sigma_{qq} \geq 0$ and $A = V\Sigma(A)W^*$. In fact,

$$\Sigma(A) = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{M}_{m,n}.$$

If A is real, V and W may be chosen to be real orthogonal.

Proof: Suppose $A^*A = W\Lambda W^*$ for a unitary $W = [W_r\ Y] \in \mathcal{M}_n$ with $W_r \in \mathcal{M}_{n,r}$, $\Lambda = \Lambda_r \oplus 0_{n-r}$, and $\Lambda_r \equiv \operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Set $\Sigma_r \equiv \Lambda_r^{\frac{1}{2}}$, $D \equiv \Sigma_r \oplus I_{n-r}$, and $X \equiv AWD^{-1}$; partition $X = [V_r\ Z]$ with $V_r \in \mathcal{M}_{m,r}$. Then

$$\begin{bmatrix} V_r^* V_r & V_r^* Z \\ Z^* V_r & Z^* Z \end{bmatrix} = X^* X = \begin{bmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{bmatrix},$$

so $Z = 0$ and $V_r^* V_r = I_r$. Moreover, $A = XDW^* = [V_r\ 0](\Sigma_r \oplus I_{n-r})[W_r\ Y]^* = V_r \Sigma_r W_r^*$, as desired. If $V = [V_r\ U] \in \mathcal{M}_m$ is unitary, then $A = XDW^* = V\Sigma(A)W^*$, as asserted. Since $A^*A = W\Sigma(A)^T\Sigma(A)W^*$, uniqueness of the nonnegative diagonal matrix $\Sigma(A)$ follows from identification of its leading q-by-q principal submatrix with the square root of the leading q-by-q principal submatrix of Λ. ∎

The diagonal entries of $\Sigma(A)$ are the *singular values* of A, and they are uniquely determined as the square roots of the q largest eigenvalues of A^*A or AA^* (the two sets of eigenvalues are the same except for $|m - n|$ zeroes [12, Theorem 1.3.20]). The *multiplicity* of a singular value of A is the number of times it appears as a diagonal entry of $\Sigma(A)$. The *rank* of A is evidently the number of its positive singular values.

The unitary factors V and W in a singular value decomposition of A are never unique. L. Autonne discovered in 1915 that their degree of non-uniqueness depends on the multiplicities of the singular values of A: Let $s_1 > \cdots > s_k > 0$ be the distinct positive singular values of A, with respective multiplicities $n_1, \ldots, n_k$, so $r = n_1 + \cdots + n_k$, and let $W, \hat{W} \in \mathcal{M}_m$ and $V, \hat{V} \in \mathcal{M}_n$ be unitary. Then $V\Sigma(A)W^* = \hat{V}\Sigma(A)\hat{W}^*$ if and only if there are unitary matrices $U_i \in \mathcal{M}_{n_i}$, $i = 1, \ldots, k$, a unitary $\tilde{V} \in \mathcal{M}_{m-r}$, and a unitary $\tilde{W} \in \mathcal{M}_{n-r}$ such that $\hat{V} = V(U_1 \oplus \cdots \oplus U_k \oplus \tilde{V})$ and $\hat{W} = W(U_1 \oplus \cdots \oplus U_k \oplus \tilde{W})$. For a detailed discussion, see [3, 19°] or [13, Theorem 3.1.1′]; the key observation is that $\hat{V}^* V$ commutes with the diagonal matrix $\Sigma(A)\Sigma(A)^T$ (thought of as a direct sum of distinct scalar multiples of identity matrices) and $\hat{W}^* W$ commutes with $\Sigma(A)^T\Sigma(A)$, so both unitary products must be block diagonal. In particular, if A is real and has n distinct singular values, and if the signs of any one nonzero entry in each column of the real orthogonal matrix V are specified, then all three real factors in the singular value decomposition of A are uniquely determined.

We have just seen that the spectral theorem for positive semidefinite Hermitian matrices (and something like the Gram-Schmidt process that permits extension of an orthonormal set to an orthonormal basis) implies the singular value decomposition. The converse is also true: If $A \in \mathcal{M}_n$ is Hermitian and positive semidefinite, and if $A = V\Sigma W^*$ is a given singular value decomposition, then V^*AV is Hermitian, so $I_n\Sigma(W^*V) = (V^*W)\Sigma I_n$ are two singular value decompositions of

V^*AV. Autonne's uniqueness theorem tells us that V^*W is a direct sum of unitary blocks whose sizes are the multiplicities of the singular values of A. Since a Hermitian positive semidefinite unitary matrix can only be an identity matrix, the blocks of V^*W corresponding to positive singular values of A are identity matrices; the diagonal block corresponding to a zero singular value is arbitrary, so we may take it to be an identity matrix. The conclusion is that the two unitary factors in the singular value decomposition of a Hermitian positive semidefinite matrix may be chosen to be the same, which gives the spectral theorem for a Hermitian positive semidefinite matrix. The spectral theorem for a general Hermitian matrix A follows from the positive semidefinite case by considering $A + \lambda I$ for some positive λ larger than the absolute value of the algebraically smallest eigenvalue of A. For another discussion of the equivalence of the singular value decomposition and the spectral theorem, and a proof of the singular value decomposition that is independent of the spectral theorem, see [10].

The history of the singular value decomposition has been documented in [13, Section 3.0] and [25]. Proofs for real square matrices were offered independently by Beltrami [4] in 1873 and by Jordan [15] in 1874. A complete singular value decomposition for square complex matrices (including a characterization of the uniqueness of the factors) was obtained in 1915 by Autonne [3], who also showed that various special factorizations for complex symmetric, skew-symmetric, orthogonal, normal, Hermitian, positive semidefinite, and Lorenzian matrices (and their real and normal versions) could all be derived from the singular value decomposition. Although the rectangular complex case follows easily from the general square case [13, Problem 1 (b) in Section 3.1], Autonne did not discuss it; Eckart and Young [8] did so in 1939. Mac Duffee's citation [16, Theorem 41.6, p. 78] of what is alleged to be Autonne's version of the singular value decomposition contains an extraneous assumption of nonsingularity that was not present in Autonne's 1915 paper.

The standard recursive formulae for transforming a given sequence of independent vectors into an orthonormal sequence were stated explicitly by Erhard Schmidt in a 1907 paper [22, p. 442]. The essence of these formulae is implicit in an 1883 paper of J. P. Gram [9], which Schmidt references in a footnote.

3. A BASIC PRINCIPLE. The answer we have in mind to the question posed in the title is in the following basic principle:

Theorem 3.1. *Let n, p, and q be given positive integers with $p \le q$, and let $A \in \mathscr{M}_{p,n}$ and $B \in \mathscr{M}_{q,n}$ be given. Then $A^*A = B^*B$ if and only if $B = VA$ for some $V \in \mathscr{M}_{q,p}$ with orthonormal columns; if A and B are real, then V may be taken to be real. If* rank *$A = p$ and $B = VA$, then V is uniquely determined as $V = BA^*(AA^*)^{-1}$.*

Proof: If $B = VA$, then $B^*B = A^*V^*VA = A^*A$ if $V^*V = I$. Conversely, if $A^*A = B^*B$, use Theorem 2.1 to write $A = V_1(\Sigma_r W_r^*)$ and $B = V_2(\Sigma_r W_r^*)$ for some $V_1 \in \mathscr{M}_{p,r}$ and $V_2 = \mathscr{M}_{q,r}$ with orthonormal columns. Augmenting V_1 to size q-by-r by adding, if necessary, a block of zeroes at the bottom preserves column orthonormality and ensures that there is a unitary $U \in \mathscr{M}_q$ such that

$$V_2 = U\begin{bmatrix} V_1 \\ 0 \end{bmatrix}.$$

Writing $U = [V\ Z]$ with $V \in \mathscr{M}_{q,p}$ gives $V_2 = VV_1$ and hence $B = V_2(\Sigma_r W_r^*) = VV_1(\Sigma_r W_r^*) = VA$. If A and B are real, then W_r, V_1, and V_2 may all be chosen to

be real, and so V may also be chosen to be real. If rank $A = p$, then AA^* is nonsingular and $B = VA$ implies $BA^* = VAA^*$, so $BA^*(AA^*)^{-1} = V$. ■

The essential idea of our proof of Theorem 3.1 is employed in Williamson's treatment of the polar decomposition [29]; it was also used by Parker [21, Theorem 5], who considered square matrices ($p = q = n$). Vinograde [27, Lemma, p. 160] treated rectangular real matrices of the same size ($p = q$); using a very different approach, Hong and Horn [11, Lemma 2.2 (a)] proved the same result in the complex case. L. Autonne's 1902 discovery of an important special case of Theorem 3.1 is discussed in Section 5.

In a typical application of Theorem 3.1, one has a given matrix X, perhaps with some special structure, and uses facts about positive semidefinite matrices to factor X^*X as $X^*X = Y^*Y$, where Y has some special form. Subject to dimensional requirements, our basic principle ensures that $X = VY$ for some V with orthonormal columns. Our first two applications are based on two natural ways to factor any positive semidefinite matrix A as $A = Y^*Y$: the Cholesky factorization and the square root.

4. THE CHOLESKY AND QR DECOMPOSITIONS. Simultaneous row and column Gaussian elimination always succeeds in reducing a Hermitian positive semidefinite matrix A to a nonnegative diagonal matrix. Starting with $A_0 \equiv A$, this classical algorithm produces a sequence of matrices $A_k = R_k^* A_{k-1} R_k$, $k = 1, 2 \dots$, each of which has at least two more zero entries than its predecessor, and is congruent to its predecessor via a nonsingular upper triangular matrix R_k of a simple form. Positive semidefiniteness of each A_k ensures that any zero diagonal entry lies in a whole row and column of zero entries [12, Problem 2, Section 7.1], so no pivoting is necessary in the algorithm, which terminates in $N \le n(n-1)/2$ steps with $A_N = D = \operatorname{diag}(d_1, \dots, d_n)$ and all $d_i \ge 0$. This gives the *Cholesky decomposition*

$$A = \left[D^{\frac{1}{2}}(R_1 \cdots R_N)^{-1}\right]^* \left[D^{\frac{1}{2}}(R_1 \cdots R_N)^{-1}\right]$$

involving an upper triangular matrix $R = D^{\frac{1}{2}}(R_1 \cdots R_N)^{-1}$ and its lower triangular conjugate transpose.

Theorem 4.1. *A given $A \in \mathcal{M}_n$ is Hermitian and positive semidefinite if and only if there is an upper triangular $R = [r_{ij}] \in \mathcal{M}_n$ such that $A = R^*R$. If, in addition, A is real, then R may be chosen to be real. If A is Hermitian and positive definite, there is a unique such R with all $r_{ii} > 0$.*

Uniqueness follows from Theorem 3.1: If A is nonsingular and $A = R_1^* R_1 = R_2^* R_2$ for upper triangular $R_1, R_2 \in \mathcal{M}_n$, then $R_1 = UR_2$ for a unique unitary U. Since $U = R_1 R_2^{-1}$ is upper triangular, it must actually be diagonal. A nonnegative diagonal unitary matrix can only be an identity.

Our basic principle now gives the *QR decomposition* [12, Theorem 2.6.1] as an immediate consequence of the Cholesky decomposition.

Corollary 4.2. *Let $A \in \mathcal{M}_{m,n}$ be given with $m \ge n$. There is a $Q \in \mathcal{M}_{m,n}$ with orthonormal columns and an upper triangular $R = [r_{ij}] \in \mathcal{M}_n$ such that $A = QR$. If, in addition, A is real, then both Q and R may be chosen to be real. If* rank $A = n$,

there are unique matrices $Q \in \mathcal{M}_{m,n}$ and $R \in \mathcal{M}_n$ such that $Q^*Q = I_n$, R is upper triangular with positive diagonal entries, and $A = QR$.

Proof: Write $A^*A = R^*R$, where $R \in \mathcal{M}_n$ is upper triangular. Theorem 3.1 now ensures that $A = QR$ for some unitary Q. If A has full rank, then A^*A is nonsingular and the factor R is unique if we require that all $r_{ii} > 0$, in which case $Q = AR^{-1}$ is also uniquely determined. ■

5. THE POLAR DECOMPOSITION. The *polar decomposition* is a convenient restatement of the singular value decomposition that can be obtained as an application of Theorem 3.1.

Corollary 5.1. *Let $A \in \mathcal{M}_{m,n}$ be given with $m \geq n$. There exists a unique Hermitian positive semidefinite $P \in \mathcal{M}_n$ and some $V \in \mathcal{M}_{m,n}$ such that $V^*V = I_n$ and $A = VP$. One always has $P = (A^*A)^{\frac{1}{2}}$. If* rank $A = n$, *V is uniquely determined as $V = A(A^*A)^{-\frac{1}{2}}$. If A is real, V and P may be taken to be real.*

Proof: Write $A^*A = P^2$, where $P = (A^*A)^{\frac{1}{2}} \in \mathcal{M}_n$ is the unique positive semidefinite square root of A^*A. Since $m \geq n$, Theorem 3.1 ensures that $A = VP$ for some $V \in \mathcal{M}_{m,n}$ with orthonormal columns. The remaining assertions are straightforward. ■

In 1902, L. Autonne proved the polar decomposition for nonsingular $A \in \mathcal{M}_n$ [1, Lemme II]; the singular case is more subtle. Autonne [2] proved the following special case of Theorem 3.1 in 1903: for a given Hermitian positive semidefinite $C \in \mathcal{M}_n$, $B \in \mathcal{M}_n$ is such that $B^*B = C$ if and only if $B = UC^{\frac{1}{2}}$ for some unitary $U \in \mathcal{M}_n$. He then observed that, for any $A \in \mathcal{M}_n$, one may take $C = A^*A$ and $B = (A^*A)^{\frac{1}{2}}$ and can conclude that $A = U(A^*A)^{\frac{1}{2}}$ for some unitary U. In 1931, Wintner and Murnaghan rediscovered the polar decomposition of a nonsingular complex square matrix [30]; their paper does not cite Autonne. Apparently stimulated by Winter and Murnaghan, Williamson [29] published in 1935 a proof of the polar decomposition for not-necessarily-square complex matrices (this follows easily from the general square case; see Problem 1 (b) in Section 3.1 of [13]). Although Williamson cites both Autonne [1] and Wintner and Murnaghan [30] for the square nonsingular case, he was apparently unaware of Autonne's priority (in [2]) for the singular square case as well. Moreover, C. C. MacDuffee, chronicler of all of matrix theory up to 1933, knew of Autonne's 1903 paper [2] and cited it as the source of the result that every Hermitian positive semidefinite matrix can be factored as $C = B^*B$ [16, Theorem 41.5, p. 77], but he failed to note Autonne's polar decomposition of a general square complex matrix in the same paper. Instead, MacDuffee cites Autonne's 1902 paper [1] and Wintner and Murnaghan's 1931 paper as sources for the polar decomposition of a *nonsingular* square complex matrix; he does not mention the singular case that Autonne had dispatched in 1903.

6. SIMULTANEOUS DIAGONALIZATION BY CONGRUENCE, A GENERALIZED SINGULAR VALUE DECOMPOSITION, AND THE CS DECOMPOSITION. In general, two given Hermitian matrices cannot be diagonalized simultaneously by similarity (commutativity is essential for this), but if both of them are

positive semidefinite, or if one is positive definite, then a simultaneous diagonalization by nonsingular congruence is possible.

Simultaneous diagonalization by congruence of a real symmetric positive definite B and a real symmetric A is a classical result with important applications in mechanics and statistics. Jordan [15, pp. 46–54] attributed an earlier discussion of the result to Weierstrass [28]. See [12, Problem 8 in Section 4.5] for a necessary and sufficient condition for two Hermitian matrices to be diagonalizable simultaneously by nonsingular congruence.

Theorem 6.1. *Let $A, B \in \mathscr{M}_n$ be Hermitian and positive semidefinite.*

(a) *Let* $\operatorname{rank}(A+B)=\rho$. *There is a nonsingular $S \in \mathscr{M}_n$ and n-by-n nonnegative diagonal matrices $D_A = \operatorname{diag}(\alpha_1, \dots, \alpha_n)$ and D_B such that*

$$(6.1)\quad A = S^*D_AS,\ B = S^*D_BS,\ \alpha_1 \geq \cdots \geq \alpha_n, \text{ and } D_A + D_B = I_\rho \oplus 0_{n-\rho}.$$

(b) *If B is positive definite, there is a nonsingular $S \in \mathscr{M}_n$ such that $A = S^*D_AS$ and $B = S^*S$, where the diagonal entries of the nonnegative diagonal matrix D_A are the decreasingly-ordered eigenvalues of AB^{-1}. Suppose AB^{-1} has k distinct eigenvalues $\mu_1 > \cdots > \mu_k$ with respective multiplicities $n_1, \dots, n_k$, so that $n_1 + \cdots + n_k = n$. If $S_1, S_2 \in \mathscr{M}_n$ are nonsingular, then $B = S_1^*S_1 = S_2^*S_2$ and $A = S_1^*D_AS_1 = S_2^*D_AS_2$ if and only if $S_1 = US_2$, where $U = U_1 \oplus \cdots \oplus U_k$ and each $U_i \in \mathscr{M}_{n_i}$ is unitary, $i = 1, \dots, k$.*

(c) *If $A + B = I_n$, the matrix S in* (6.1) *may be taken to be unitary and $D_A + D_B = I_n$.*

(d) *If A and B are real, then the matrix S in* (a)–(c) *may be taken to be real.*

Proof: Write $A + B = U(\Lambda_\rho \oplus 0_{n-\rho})U^*$, where $\Lambda_\rho = \operatorname{diag}(\lambda_1, \dots, \lambda_\rho)$ is nonnegative and nonsingular and $\rho \geq 1$; the case $\rho = 0$ is trivial. Let $T \equiv (\Lambda_\rho^{\frac{1}{2}} \oplus I_{n-\rho})U^*$, so $T^{-*}(A+B)T^{-1} = I_\rho \oplus 0_{n-\rho} = T^{-*}AT^{-1} + T^{-*}BT^{-1} \equiv \hat{A} + \hat{B}$. Writing this identity with conformally-partitioned block matrices gives

$$\begin{bmatrix} I_\rho & 0 \\ 0 & 0_{n-\rho} \end{bmatrix} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{12}^* & \hat{A}_{22} \end{bmatrix} + \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{12}^* & \hat{B}_{22} \end{bmatrix} \text{with } \hat{A}_{22}, \hat{B}_{22} \in \mathscr{M}_{n-\rho},$$

and hence $\hat{A}_{11} + \hat{B}_{11} = I_\rho$ and $\hat{A}_{22} + \hat{B}_{22} = 0$. Since $\hat{A}_{22}$ and $\hat{B}_{22}$ are Hermitian and positive semidefinite, they must both be zero; positive semidefiniteness of $\hat{A}$ and $\hat{B}$ then implies that $\hat{A}_{12} = \hat{B}_{12} = 0$ as well. If $V \in \mathscr{M}_\rho$ is unitary and $\hat{A}_{11} = V\Delta V^*$ for a diagonal $\Delta = \operatorname{diag}(\alpha_1, \dots, \alpha_\rho) \in \mathscr{M}_\rho(\mathbb{R})$ with $\alpha_1 \geq \cdots \geq \alpha_\rho$, then $I_\rho = V^*I_\rho V = V^*(\hat{A}_{11} + \hat{B}_{11})V = \Delta + V^*\hat{B}_{11}V$ and hence $V^*\hat{B}_{11}V = I_\rho - \Delta$ is also diagonal (and necessarily nonnegative). Finally, set $W \equiv V \oplus I_{n-\rho} \in \mathscr{M}_n$, $D_A \equiv \Delta \oplus 0_{n-\rho}$, $D_B \equiv (I_\rho - \Delta) \oplus 0_{n-\rho}$, and $S \equiv W^*T$, and observe that $A = S^*D_AS$ and $B = S^*D_BS$.

If A and B are real, the unitary matrices U and V in the preceding argument may be chosen to be real orthogonal, which makes S real.

If $A + B = I_n$, then $\rho = n$, $T = I$, and $S = W^*$ in the preceding argument.

If B is positive definite, then the eigenvalues of $AB^{-1} = (AB^{-\frac{1}{2}})B^{-\frac{1}{2}}$ are the same as those of the Hermitian positive semidefinite matrix $B^{-\frac{1}{2}}(AB^{-\frac{1}{2}})$. Write $B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = WD_AW^*$ for some unitary $W \in \mathscr{M}_n$, $D_A = \operatorname{diag}(\alpha_1, \dots, \alpha_n)$, and $\alpha_1 \geq \cdots \geq \alpha_n$. Set $S \equiv W^*B^{\frac{1}{2}}$, so that $S^*S = B^{\frac{1}{2}}WW^*B^{\frac{1}{2}} = B$ and $S^*D_AS = B^{\frac{1}{2}}(WD_AW^*)B^{\frac{1}{2}} = B^{\frac{1}{2}}(B^{-\frac{1}{2}}AB^{-\frac{1}{2}})B^{\frac{1}{2}} = A$. If A and B are real, then so are $B^{\frac{1}{2}}$ and $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, so W may be chosen to be real, in which case S is real. If $S_1^*S_1 = S_2^*S_2$,

then Theorem 3.1 ensures that $S_1 = US_2$ for a unique unitary U. If $S_1^* D_A S_1 = S_2^* D_A S_2$ as well, then $S_1^* D_A S_1 = S_2^* U^* D_A U S_2 = S_2^* D_A S_2$, so $U^* D_A U = D_A$ and $D_A U = U D_A$. Commutativity with the diagonal matrix D_A forces U to have the asserted block diagonal structure. ■

If $A, B \in \mathscr{M}_n(\mathbb{R})$ are symmetric and B is positive definite, if $A = S^T D_A S$ and $B = S^T S$ for a given nonsingular $S \in \mathscr{M}_n(\mathbb{R})$, if AB^{-1} has n distinct eigenvalues, and if the signs of any one nonzero entry in each row of S are prescribed, the observations in (b) show that S is uniquely determined.

Our basic principle now permits us to deduce from Theorem 6.1 a generalization, for a pair of matrices, of the singular value decomposition.

Corollary 6.2. *Let $A \in \mathscr{M}_{p,n}$ and $B \in \mathscr{M}_{q,n}$ be given with $q \geq n$, define $X \in \mathscr{M}_{p+q,n}$ by $X^* = [A^* \;\; B^*]$, let* rank $X = \rho$, *and let* rank $A = r$.

(*a*) *There are unitary matrices $V = [V_1 \;\; V_2] \in \mathscr{M}_p$ and $W = [W_1 \;\; W_2] \in \mathscr{M}_q$ with $V_1 \in \mathscr{M}_{p,r}$ and $W_1 \in \mathscr{M}_{q,n}$, nonnegative diagonal matrices $\Delta_B \in \mathscr{M}_n$ and $D_r = \operatorname{diag}(\delta_1, \ldots, \delta_r)$, and a nonsingular $S \in \mathscr{M}_n$ such that $\delta_1 \geq \cdots \geq \delta_r > 0$, $(D_r^2 \oplus 0_{n-r}) + \Delta_B^2 = I_\rho \oplus 0_{n-\rho}$,*

$$A = V_1[D_r \;\; 0]S = V\Sigma_A S, \text{ and } B = W_1 \Delta_B S = W \Sigma_B S, \tag{6.2}$$

where

$$\Sigma_A = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathscr{M}_{p,n} \text{ and } \Sigma_B = \begin{bmatrix} \Delta_B \\ 0 \end{bmatrix} \in \mathscr{M}_{q,n}. \tag{6.3}$$

If A and B are real, then V, W, and S may all be chosen to be real.

(*b*) *Suppose* rank $B = n$. *Then one may take $\Delta_B = I_n$ in* (6.2) *and* (6.3), *in which case the diagonal entries of Σ_A are the decreasingly-ordered square roots of the eigenvalues of $\Psi \equiv (A^*A)(B^*B)^{-1}$. Suppose Ψ has k distinct positive eigenvalues $\mu_1 > \cdots > \mu_k > 0$ with respective multiplicities $n_1, \ldots, n_k$ so that $n_1 + \cdots + n_k = r$; set $n_{k+1} \equiv n - r$. Then nonsingular matrices $S_1, S_2 \in \mathscr{M}_n$ can serve as S-factors in* (6.2) *if and only if $S_1 = US_2$, where $U = U_1 \oplus \cdots \oplus U_k$ and each $U_i \in \mathscr{M}_{n_i}$ is unitary, $i = 1, \ldots, k+1$. Given S, the factor $W_1 = BS^{-1}$ in* (6.2) *is unique; if, in addition,* rank $A = p$, *then the factor V_1 is also unique.*

(*c*) *If $p = n$ and $A^*A + B^*B = I_p$, then S may be taken to be unitary in* (6.2) *and $\Sigma_A^2 + \Delta_B^2 = I_p$.*

Proof: Apply Theorem 6.1 to the two positive semidefinite Hermitian matrices A^*A and B^*B (for which rank X = rank X^*X = rank$(A^*A + B^*B) = \rho$) and conclude that there is a nonsingular $S \in \mathscr{M}_n$ and n-by-n nonnegative diagonal matrices $D_A = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ and D_B such that $\alpha_1 \geq \cdots \geq \alpha_r > \alpha_{r+1} = \cdots = \alpha_n = 0$, and $D_A + D_B = I_\rho \oplus 0_{n-\rho}$. Write $D_A^{\frac{1}{2}} \equiv D_r \oplus 0_{n-r}$ and $D_B^{\frac{1}{2}} \equiv \Delta_B$, and define $\Sigma_A \in \mathscr{M}_{p,n}$ and $\Sigma_B \in \mathscr{M}_{q,n}$ by (6.3) (here we use the hypothesis that $q \geq n$). Note that $D_A = \Sigma_A^T \Sigma_A$, $A^*A = (\Sigma_A S)^*(\Sigma_A S)$, and $B^*B = (\Delta_B S)^*(\Delta_B S)$. Theorem 3.1 ensures that there are unitary matrices $V \in \mathscr{M}_p$ and $W \in \mathscr{M}_q$ such that $A = V\Sigma_A S$ and $B = W\Sigma_B S$. Writing $V = [V_1 \;\; V_2]$ and $W = [W_1 \;\; W_2]$ with $V_1 \in \mathscr{M}_{p,r}$ and $W_1 \in \mathscr{M}_{q,n}$ gives the alternate representations in (6.2). ■

If B is square and nonsingular, the diagonal entries of D_r in (6.3) are the r positive singular values of AB^{-1}. If $B = I_n$, then $\Sigma_B = I_n$ and $S = W^*$, so $A = V\Sigma_A W^*$ in (6.2) and we recover the ordinary singular value decomposition in Theorem 2.1.

If $A, B \in \mathscr{M}_n$ are real and nonsingular and all the singular values of AB^{-1} are distinct, then all the factors in the representations $A = V\Sigma_A S$ and $B = WS$ (V and W are real orthogonal and S is nonsingular) are uniquely determined if the signs of any one nonzero entry in each row of S are prescribed.

The idea in Corollary 6.2(c) yields the following useful and fundamental factorization—the *CS Decomposition*—for a partitioned unitary matrix.

Corollary 6.3. *Let p and q be given positive integers with $p \le q$, let $U \in \mathscr{M}_{p+q}$ be unitary, and partition U as*

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \text{ with } U_{11} \in \mathscr{M}_p \text{ and } U_{22} \in \mathscr{M}_q.$$

Let the diagonal entries of the diagonal matrix $C \in \mathscr{M}_p$ be the decreasingly-ordered singular values of U_{11}, all of which lie in the interval $[0, 1]$, and define $S \equiv (I_p - C)^{\frac{1}{2}}$, so $C^2 + S^2 = I_p$. There are unitary matrices $V_1, W_1 \in \mathscr{M}_p$ and $V_2, W_2 \in \mathscr{M}_q$ such that

$$U = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \begin{bmatrix} C & -S & 0 \\ S & C & 0 \\ 0 & 0 & I_{q-p} \end{bmatrix} \begin{bmatrix} W_1^* & 0 \\ 0 & W_2^* \end{bmatrix}, \tag{6.4}$$

that is, $U_{11} = V_1 C W_1^*$, $U_{12} = V_1[-S \ 0_{p,q-p}]W_2^*$, $U_{21}^* = W_1[S \ 0_{p,q-p}]V_2^*$, and $U_{22} = V_2(C \oplus I_{q-p})W_2^*$. In particular, the singular values of U_{11}, U_{12}, U_{21}, and U_{22} are the diagonal entries of C, S, S, and $C \oplus I_{q-p}$, respectively. If U is real, then V_1, V_2, W_1, and W_2 may all be chosen to be real orthogonal.

Proof: Apply Corollary 6.2 (c) with $n = p$, $A = U_{11}$, and $B = U_{21}$. Conclude that

$$\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \begin{bmatrix} C \\ S \\ 0_{q-p} \end{bmatrix} W_1^*$$

for some unitary $V_1, W_1 \in \mathscr{M}_p$ and $V_2 \in \mathscr{M}_q$, where $C \equiv \Sigma_A$ and $S \equiv \Delta_B = (I_p - \Sigma_A^2)^{\frac{1}{2}}$. Since the unitary matrix $V_1 \oplus V_2$ maps the column space of

$$\begin{bmatrix} C \\ S \\ 0_{q-p} \end{bmatrix} W_1^*$$

onto the column space of

$$\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix},$$

it must also map the orthogonal complement of the former subspace onto the orthogonal complement of the latter. The respective orthogonal complements are easily verified to be the column spaces of

$$\begin{bmatrix} -S & 0 \\ C & 0 \\ 0 & I_{q-p} \end{bmatrix} \begin{bmatrix} W_1^* & 0 \\ 0 & I_{q-p} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix}.$$

Thus, there is a unitary $Z \in \mathcal{M}_q$ such that

$$\begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \begin{bmatrix} -S & 0 \\ C & 0 \\ 0 & I_{q-p} \end{bmatrix} \begin{bmatrix} W_1^* & 0 \\ 0 & I_{q-p} \end{bmatrix} = \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix} Z.$$

Setting $W_2 \equiv Z(W_1 \oplus I_{q-p})$ gives (6.4). ∎

There are many useful generalizations of the singular value decomposition to pairs of matrices, some motivated by applications to constrained generalized least squares problems. The essence of Corollary 6.2 (a), (b) (for certain real square matrices A and B with full rank) is in Olkin's 1951 dissertation [17] and in [18], where it was used to find the joint distribution of the singular values of a product AB^{-1}, for random matrices A and B. The factorization (6.2) is called the *B-singular value decomposition* in [26]. Other generalizations are in [19] and [7].

The ideas involved in the CS decomposition (6.4) were introduced by Davis and Kahan in [5] and [6]. According to the comprehensive history in [20], the first complete statement and proof of the CS decomposition were in [23]; Stewart coined the name, which first appeared in print in [24]. See [20] for many references and applications.

7. CANONICAL CORRELATIONS AND STRUCTURED FACTORIZATIONS OF A PARTITIONED MATRIX. Let $A \in \mathcal{M}_{p+q}$ be Hermitian positive semidefinite and partitioned as a 2-by-2 block matrix

$$\mathscr{A} = \begin{bmatrix} \mathscr{A}_{11} & \mathscr{A}_{12} \\ \mathscr{A}_{12}^* & \mathscr{A}_{22} \end{bmatrix}, \mathscr{A}_{11} \in \mathcal{M}_p, \mathscr{A}_{22} \in \mathcal{M}_q, q \geq p. \tag{7.1}$$

We want to reduce $\mathscr{A}$ to a canonical form with a block diagonal congruence $\mathscr{A} \to L^*\mathscr{A}L$, $L = F \oplus G$, $F \in \mathcal{M}_p$, $G \in \mathcal{M}_q$, sometimes called an *inner congruence*.

Because $\mathscr{A}_{11}$ and $\mathscr{A}_{22}$ are positive semidefinite, there are unitary $U_1 \in \mathcal{M}_p$ and $U_2 \in \mathcal{M}_q$ such that $U_i^*\mathscr{A}_{ii}U_i = \Lambda_i \oplus 0$, where $\Lambda_i \in \mathcal{M}_{r_i}$ is nonsingular and nonnegative diagonal and $r_i = \operatorname{rank} \mathscr{A}_{ii}$. Let $U \equiv U_1 \oplus U_2$. Since $U^*\mathscr{A}U$ is Hermitian and positive semidefinite, its diagonal zero blocks must lie in entire rows and columns of zeroes:

$$U^*\mathscr{A}U = \begin{bmatrix} \Lambda_1 & 0 & \mathscr{B}_{12} & 0 \\ 0 & 0_{p-r_1} & 0 & 0 \\ \mathscr{B}_{12}^* & 0 & \Lambda_2 & 0 \\ 0 & 0 & 0 & 0_{q-r_2} \end{bmatrix}, \qquad \mathscr{B}_{12} \in \mathcal{M}_{r_1, r_2}.$$

Now let $S \equiv (\Lambda_1^{-\frac{1}{2}} \oplus I_{p-r_1}) \oplus (\Lambda_2^{-\frac{1}{2}} \oplus I_{q-r_2})$, so

$$S^* U^*\mathscr{A}US = \begin{bmatrix} I_{r_1} & 0 & \mathscr{C}_{12} & 0 \\ 0 & 0_{p-r_1} & 0 & 0 \\ \mathscr{C}_{12}^* & 0 & I_{r_2} & 0 \\ 0 & 0 & 0 & 0_{q-r_2} \end{bmatrix}, \qquad \mathscr{C}_{12} \in \mathcal{M}_{r_1, r_2}.$$

Finally, let $\mathscr{C}_{12} = V\Sigma W^*$ be a singular value decomposition, with unitary $V \in \mathscr{M}_{r_1}$ and $W \in \mathscr{M}_{r_2}$, and a nonnegative diagonal $\Sigma \in \mathscr{M}_{r_1, r_2}$ with decreasingly-ordered diagonal entries. Setting $T \equiv (V \oplus I_{p-r_1}) \oplus (W \oplus I_{q-r_2})$, we have

$$T^*S^*U^*\mathscr{A}(UST) = \begin{bmatrix} I_{r_1} & 0 & \Sigma & 0 \\ 0 & 0_{p-r_1} & 0 & 0_{p-r_1, q-r_2} \\ \Sigma^T & 0 & I_{r_2} & 0 \\ 0 & 0 & 0 & 0_{q-r_2} \end{bmatrix}. \tag{7.2}$$

The diagonal entries of Σ are the $\min\{r_1, r_2\}$ largest singular values of $(\mathscr{A}_{11}^{\dagger})^{\frac{1}{2}}\mathscr{A}_{12}(\mathscr{A}_{22}^{\dagger})^{\frac{1}{2}}$.

Now write $E_1 \equiv I_{r_1} \oplus 0_{p-r_1}$, define $E_2 \in \mathscr{M}_p$ and $E_3 \in \mathscr{M}_{q-p}$ by $E_2 \oplus E_3 = I_{r_2} \oplus 0_{q-r_2}$, define $D \equiv \operatorname{diag}(\delta_1, \dots, \delta_p) \in \mathscr{M}_p$ by

$$\begin{bmatrix} \Sigma & 0 \\ 0 & 0_{p-r_1, q-r_2} \end{bmatrix} = \begin{bmatrix} D & 0_{p, q-p} \end{bmatrix}, \quad \delta_1 \geq \cdots \geq \delta_p \geq 0,$$

and re-write (7.2) in terms of $L \equiv US^{-1}T = F \oplus G (F \in \mathscr{M}_p$ and $G \in \mathscr{M}_q)$ as

$$\mathscr{A} = L \begin{bmatrix} E_1 & D & 0 \\ D & E_2 & 0 \\ 0 & 0 & E_3 \end{bmatrix} L^* \equiv L\mathscr{D}L^*. \tag{7.3}$$

This is the desired canonical form. Positive semidefiniteness of 2-by-2 principal submatrices of $\mathscr{D}$ implies that the diagonal entries of E_1, E_2, and D satisfy the inequalities $(E_1)_{ii}(E_2)_{ii} \geq \delta_i^2$, $i = 1, \dots, p$, so all $\delta_i \in [0, 1]$ and δ_i can be nonzero only for $i = 1, \dots, \operatorname{rank} \mathscr{A}_{12} \leq \min\{\operatorname{rank} \mathscr{A}_{11}, \operatorname{rank} \mathscr{A}_{22}\} \leq p$. Expressing (7.3) as an inner congruence, we have shown that there are nonsingular $F \in \mathscr{M}_p$ and $G \in \mathscr{M}_q$ such that

$$\mathscr{A}_{11} = FE_1F^*, \mathscr{A}_{12} = F[D \quad 0]G^*, \mathscr{A}_{22} = G(E_2 \oplus E_3)G^*. \tag{7.4}$$

If $\mathscr{A}_{11}$ and $\mathscr{A}_{22}$ are nonsingular (which is the case if $\mathscr{A}$ is positive definite), then $E_1 = I_p$, $E_2 \oplus E_3 = I_q$, and the diagonal entries of D are the decreasingly-ordered nonnegative square roots of the eigenvalues of $\mathscr{A}_{11}^{-1}\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\mathscr{A}_{12}^*$.

Hotelling's seminal paper [14] on the relations between two sets of variates provided a generalization of the multiple correlation. He discovered the factorization (7.4) in 1935 for a real symmetric positive definite $\mathscr{A}$, in which case the diagonal entries of D are known as *canonical correlations*. This model is a basic multivariate statistical analysis now known as *canonical analysis*.

In order to use this canonical form with our basic principle, we need to express the block matrix in (7.3) as $\mathscr{D} = Y^*Y$ for some useful Y. Two natural choices are readily verified: A Cholesky decomposition gives $\mathscr{D} = R_1^*R_1$ with

$$R_1 \equiv \begin{bmatrix} E_1 & D & 0 \\ 0 & (E_2 - D^2)^{\frac{1}{2}} & 0 \\ 0 & 0 & E_3 \end{bmatrix}, \tag{7.5}$$

and a square root gives $\mathscr{D} = R_2^* R_2 = R_2^2$, with

$$R_2 \equiv \begin{bmatrix} C_r \oplus K_1 & S_r \oplus 0_{p-r} & 0 \\ S_r \oplus 0_{p-r} & C_r \oplus K_2 & 0 \\ 0 & 0 & E_3 \end{bmatrix}, \tag{7.6}$$

$r = \min\{r_1, r_2\}$, K_1 and K_2 defined by $E_1 = I_r \oplus K_1$ and $E_2 = I_r \oplus K_2$, $D_r \in \mathscr{M}_r$ defined by $D = D_r \oplus 0_{p-r}$, $C_r \equiv \frac{1}{2}[I_r + (I_r - D_r^2)^{\frac{1}{2}}]^{\frac{1}{2}}$, and $S_r \equiv \frac{1}{2}[I_r - (I_r - D_r^2)^{\frac{1}{2}}]^{\frac{1}{2}}$. Notice that at least one of K_1, K_2 is a zero matrix and that $C_r^2 + S_r^2 = I_r$.

Of course, we have in mind a positive semidefinite block matrix $\mathscr{A}$ of a special form. For given matrices A and B with the same number of columns, let $X^* = [A^* \;\; B^*]$ and consider $\mathscr{A} = XX^*$. Subject to dimensional requirements, Theorem 3.1 and the preceding considerations permit us to conclude that $X = VR_1L^* = WR_2L^*$ for some V, W with orthonormal columns. This gives the following structured factorizations for A and B:

Corollary 7.1. *Let $A \in \mathscr{M}_{p,n}$ and $B \in \mathscr{M}_{q,n}$ be given with $q \geq p$ and $n \geq p + q$. Let $r_1 = \operatorname{rank} A$, $r_2 = \operatorname{rank} B$, and $r = \min\{r_1, r_2\}$. There are nonsingular $F \in \mathscr{M}_p$ and $G \in \mathscr{M}_q$, and matrices $V, W \in \mathscr{M}_{n,p+q}$ with orthonormal columns such that*

$$A = F\begin{bmatrix} E_1 & 0_p & 0_{p,q-p} \end{bmatrix}V^* \quad \textit{and} \quad B = G\begin{bmatrix} D & (E_2 - D^2)^{\frac{1}{2}} & 0 \\ 0 & 0 & E_3 \end{bmatrix}V^* \tag{7.7}$$

and

$$\begin{aligned} A &= F\begin{bmatrix} C_r \oplus K_1 & S_r \oplus 0_{p-r} & 0_{p,q-p} \end{bmatrix}W^* \quad \textit{and} \\ B &= G\begin{bmatrix} S_r \oplus 0_{p-r} & C_r \oplus K_2 & 0 \\ 0 & 0 & E_2 \end{bmatrix}W^*, \end{aligned} \tag{7.8}$$

where $E_1, E_2, D \in \mathscr{M}_p$, $E_3 \in \mathscr{M}_{q-p}$, $C_r, S_r \in \mathscr{M}_r$, and $K_1, K_2 \in \mathscr{M}_{p-r}$ are nonnegative diagonal matrices defined as follows: $E_1 = I_{r_1} \oplus 0_{p-r_1}$, $E_2 \oplus E_3 = I_{r_2} \oplus 0_{q-r_2}$, the diagonal entries of D are the decreasingly-ordered singular values of $((AA^*)^\dagger)^{\frac{1}{2}}(AB^*)((BB^*)^\dagger)^{\frac{1}{2}}$, $E_1 = I_r \oplus K_1$, $E_2 = I_r \oplus K_2$, $D = D_r \oplus 0_{p-r}$, $C_r = \frac{1}{2}\left[I_r + \left(I_r - D_r^2\right)^{\frac{1}{2}}\right]^{\frac{1}{2}}$, and $S_r = \frac{1}{2}\left[I_r - \left(I_r - D_r^2\right)^{\frac{1}{2}}\right]^{\frac{1}{2}}$. If A and B are real, then F, G, and W may all be chosen to be real.

The factorization (7.8) (for certain real matrices A and B with full rank) is in [17].

8. THE MATRICES A^*A, $\bar{A}A$, and AA^*. Although A^*A, $\bar{A}A$, and AA^* are all the same for complex scalars, the example $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ shows that they may all be different for square matrices. The matrices

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

illustrate that it is possible to have $A^*A = B^*B$ and $\bar{A}A = \bar{B}B$ without also having $AA^* = BB^*$. However, if $A, B \in \mathcal{M}_n$ are nonsingular and $A^*A = B^*B$, then there is some unitary $U \in \mathcal{M}_n$ such that $B = UA$. If $\bar{A}A = \bar{B}B$ as well, then $\bar{A}A = \bar{U}\bar{A}UA$, so $\bar{A} = \bar{U}\bar{A}U$ and $AU^T = UA$. Thus, $B = UA = AU^T$, so $BB^* = AA^*$. This proves the following corollary of Theorem 3.1, which arose in [11] in a study of the unitary congruence $A = UBU^T$.

Corollary 8.1. *Let $A, B \in \mathcal{M}_n$ be nonsingular. If $A^*A = B^*B$ and $\bar{A}A = \bar{B}B$, then $AA^* = BB^*$ and there is a unitary $U \in \mathcal{M}_n$ such that $B = UA = AU^T$.*

REFERENCES

1. L. Autonne, Sur les groupes linéaires, réels et orthogonaux, *Bull. Soc. Math. France* **30** (1902), 121–134.
2. L. Autonne, Sur l'hypohermitien, *Bull. Soc. Math. France* **31** (1903), 140–155.
3. L. Autonne, Sur les matrices hypohermitiennes et sur les matrices unitaires, Annales de l'Université de Lyons, Nouvelle Série, Fasc. **38** (1915), 1–77.
4. E. Beltrami, Sulle funzioni bilineari, *Giornale de Mathematiche* **11** (1873), 98–106.
5. C. Davis and W. M. Kahan, Some new bounds on perturbations of subspaces, *Bull. Amer. Math. Soc.* **75** (1969), 863–868.
6. C. Davis and W. M. Kahan, The rotation of eigenvectors by a perturbation III, *SIAM J. Numer. Analysis* **7** (1970), 1–46.
7. B. De Moor and H. Zha, A tree of generalizations of the ordinary singular value decomposition, *Linear Algebra Appl.* **147** (1991), 469–500.
8. C. Eckart and G. Young, A principal axis transformation for non-Hermitian matrices, *Bull. Amer. Math. Soc.* **45** (1939), 118–121.
9. J. P. Gram, Ueber die Entwickelung reeler Functionen in Reihen mittelst der Methode der kleinsten Quadrate, *J. für reine angew. Math.* (*Crelle's Journal*) **94** (1883), 41–73.
10. K. Hoecksmann, Singular values and the spectral theorem, *Amer. Math. Monthly* **97** (1990), 413–414.
11. Y. Hong and R. A. Horn, A characterization of unitary congruence, *Linear Multilinear Algebra* **25** (1989), 105–119.
12. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
13. R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.
14. H. Hotelling, Relations between two sets of variates, *Biometrika* **28** (1936), 321–377.
15. C. Jordan, Mémoire sur les formes bilinéaires, *J. Math. Pures Appl.* (2) **19** (1874), 35–54.
16. C. C. MacDuffee, *The Theory of Matrices*, J. Springer, Berlin, 1933; reprinted by Chelsea Publishing Company, New York, 1946.
17. I. Olkin, *On Distribution Problems in Multivariate Analysis*, Doctoral Dissertation, University of North Carolina, 1951.
18. I. Olkin and S. N. Roy, On multivariate distribution theory, *Ann. Math. Statist.* **25** (1954), 329–339.
19. C. C. Paige and M. A. Saunders, Towards a generalized value decomposition, *SIAM J. Numer. Anal.* **18** (1981), 398–405.
20. C. C. Paige and M. Wei, History and generality of the CS decomposition, *Linear Algebra Appl.* 208/209 (1994), 303–326.
21. W. V. Parker, The characteristic roots of matrices, *Duke Math. J.* **12** (1945), 519–526.
22. E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen, *Math. Annalen* **63** (1907), 433–476.
23. G. W. Stewart, On the perturbation of pseudo-inverses, projections, and linear least squares problems, *SIAM Review* **19** (1977), 634–662.
24. G. W. Stewart, Computing the CS decomposition of a partitioned orthogonal matrix, *Numer. Math.* **40** (1982), 297–306.
25. G. W. Stewart, On the early history of the singular value decomposition, *SIAM Review* **35** (1993), 551–566.

26. C. VanLoan, Generalizing the singular value decomposition, *SIAM J. Numer. Anal.* **13** (1976), 76–83.
27. B. Vinograde, Canonical positive definite matrices under internal linear transformations, *Proc. Amer. Math. Soc.* **1** (1950), 159–161.
28. K. Weierstrass, Zur Theorie der bilinearen und quadratischen Formen, *Monatsber. Akad. Wiss. Berlin* (1868), 310–338.
29. J. Williamson, A polar representation of singular matrices, *Bull. Amer. Math. Soc.* **41** (1935), 118–123.
30. A. Wintner and F. D. Murnaghan, On a polar representation of non-singular square matrices, *Proc. National Acad. Sciences USA* **17** (1931), 676–678.

Mathematics Department
University of Utah
Salt Lake City, UT 84112
rhorn@math.utah.edu

Statistics Department
Stanford University
Stanford, CA 94305
iolkin@playfair.stanford.edu

PICTURE PUZZLE

(from the collection of Paul Halmos)

Two friends and collaborators.
(see page 504)

Our final observation concerns the first open problem in **[1]** and Theorem 1 in **[4]**. It seems improbable to characterize sets $A \subset \mathbb{N}$ for which $R(A)$ is dense in $\mathbb{R}^+$ by conditions involving limits as in **[4]** Theorem 1.

Indeed, as the subsequent example demonstrates, if we are given an infinite set $A \subset \mathbb{N}$, it is sufficient to add (in comparison with A) only a "few" elements to A and the resulting set will already be dense in $\mathbb{R}^+$. More precisely, if the elements of A do not satisfy some equality involving limits, then the resulting set will not satisfy it either; however, its ratio set will be dense in $\mathbb{R}^+$.

Example. Let $a_1 < a_2 < \cdots$ be an arbitrary infinite sequence of positive integers. Choose an arbitrary subsequence $(a_{n_k})_{k=1}^{\infty}$ of $(a_n)_{n=1}^{\infty}$. Denote by $(r_n)_{n=1}^{\infty}$ the sequence containing every rational number greater than 1 infinitely many times. Let $b_k = \lfloor a_{n_k} \cdot r_k \rfloor$ for $k = 1, 2 \dots$. We will show that the ratio set of $C = \{a_k; k \in \mathbb{N}\} \cup \{b_k; k \in \mathbb{N}\}$ is dense in $\mathbb{R}^+$. It suffices to show that $\mathbb{Q}^+ \subset R(C)^d$.

Let $s \in \mathbb{Q}^+$. Without loss of generality, we may assume that $s > 1$. In view of the definition of $(r_n)_{n=1}^{\infty}$, we get a sequence $(m_l)_{l=1}^{\infty}$ of positive integers such that $r_{m_l} = s$ for all $l \in \mathbb{N}$. Then

$$\frac{b_{m_l}}{a_{n_{m_l}}} = \frac{\lfloor a_{n_{m_l}} \cdot s \rfloor}{a_{n_{m_l}}} \to s \text{ as } l \to \infty;$$

thus, $s \in R(C)^d$.

REFERENCES

1. D. Hobby, D. M. Silberger, Quotients of primes, *Amer. Math. Monthly* **100** (1993), 50–52.
2. W. Narkiewicz, T. Šalát, A theorem of H. Steinhaus and (R)-dense sets of positive integers, *Czechoslov. Math. J.* **34** (109) (1984), 355–361.
3. T. Šalát, On ratio sets of natural numbers, *Acta Arith.* **15** (1969), 273–278.
4. P. Starni, Answers to two questions concerning quotients of primes, *Amer. Math. Monthly* **102** (1995), 347–349.
5. J. T. Tóth, L. Zsilinszky, On density of ratio sets of powers of primes, *Nieuw Archief voor Wiskunde* **13** (1995), 205–208.

Department of Mathematics
University of Education
Farská 19, 949 74 Nitra, Slovakia
e-mail: bukor@unitra.sk, toth@unitra.sk

Answer to Picture Puzzle

(p. 482)

Kennan T. Smith and William F. Donoghue in 1969

A Sampler of Randomness

Susan Bassein

In his note [7], Kac asked the question "What is random?" and presented evidence that at the very least, there is no consensus on an answer, and that very possibly there is no absolute answer. In order to explore that question, this paper samples concepts of randomness from information theory, number theory, complexity theory, and physics. Further, it shows how to illustrate some of these ideas using electronic devices that are small, inexpensive, and simple to build.

One of the troublesome points in answering Kac's question can be addressed by distinguishing between applying the word "random" to a finite sequence of numbers as opposed to a process that generates a sequence of numbers. The obvious point here is that flipping a coin n times, a process that is generally considered a paradigm of a random process, produces any of the 2^n possible sequences of heads or tails with equal likelihood; in that sense, then, a sequence of n heads is just as random as any other sequence.

Let us consider the issue of the randomness of a sequence of numbers first. The intuitive criterion for such a sequence to be considered random—technically, *pseudo-random*—is that it "look random", in other words, that it exhibit *disorder.* Knuth [8] presents various statistical tests for measuring the disorder of a finite sequence of numbers, the simplest of which is, for example, determining the relative frequency of all possible patterns of length 1, 2, and so on. Chaitin [4] and Kolmogorov [9] take a different approach as to whether a sequence "looks random". In their algorithmic information theory, they measure the randomness of a (long) finite sequence of bits by the length of the shortest computer program that can produce that sequence of bits: if the program is substantially shorter than the sequence, then the sequence is not random. (To make this precise, Chaitin [5] defines exactly what he means by a program.) For example, a sequence of n 1's can be generated by the program:

1. Set i equal to 1.
2. Print "1".
3. If $i = n$, then stop.
4. Add 1 to i.
5. Go back to Step 2.

No matter how large n is, this program takes only a small, fixed number of bits more than the representation of the integers i and n, so its length is essentially $2\log_2 n$ (on a binary computer). Since that will be much less than n if n is large, the sequence is not random.

A sequence of alternating 1's and 0's will lengthen the program only slightly, so that's still not random, according to this idea. But if the sequence is disorderly enough, then any program that prints it out will not be much shorter than the sequence itself and the sequence will be random. Therefore, to prove that a

sequence is random, one must show that all programs substantially shorter than the sequence cannot produce the sequence. By the usual sort of counting argument, if n is large, then most sequences of length n are random since there will be many fewer programs whose length is substantially less than n. However, Chaitin uses the undecidability of the halting program to show that if a *given* sequence is so long that its complexity is greater than that of the system of arithmetic, then it is generally impossible to prove that it is random.

Approaches to randomness that are based on the process instead of the sequence rely more on the concept of *unpredictability*, rather than *disorder*: in this context, "looking random" means that the next number produced is unpredictable. Then we must ask, "Unpredictable by whom?" To elaborate on that point, we will consider a sequence of examples of increasing unpredictability, which corresponds to decreasing populations of predictors. Before doing that, we can recast the Chaitin–Kolmogorov concept of a random sequence in the language of unpredictability of a process. One way is to consider a random process such as coin flipping. Then we can say that, by the same counting argument referred to above, the probability is extremely close to 1 that a sequence of heads and tails produced by that process will be a random sequence of length n. Or, turning it around the other way, given a random sequence, we can say that no reasonably short program could predict the last bit from the first $n - 1$ bits. From this point of view, that last bit is not predictable by anyone.

If we relax the standard of absolute unpredictability a little to allow one agent —the generator of the sequence—to be able to predict, i.e., compute, the next number in the sequence, then we have a pseudo-random number generator. This can be viewed as replacing the question "Is there a reasonably fast algorithm for predicting the next number?" by the question "Can anyone (but the generator) find, in a reasonably short amount of time, a reasonably fast algorithm for predicting the next number?" Blum, Blum, and Shub [3], in presenting algorithms for pseudo-random number generators, ask how difficult it is for someone with knowledge of a portion of the sequence to predict the next number. They use a standard measure of complexity to measure that difficulty: the process is not random, i.e., not unpredictable, if there is an algorithm whose running time is a polynomial in some reasonably small parameter of the generator and that correctly predicts the next number significantly more often than random guessing would.

The first algorithm they examine is the $1/P$ generator: given a prime P, and a number-base $1 < b < P$ that is a primitive root mod P, simply generate the sequence of digits in the base-b representation of $1/P$. (If you don't start from the first digit of that representation, it is equivalent to choosing some $0 < r < P$ and using r/P instead.) Although they show that the sequence "looks random" because its period is $P - 1$ (and P might be a number with, say, 50 or more digits) and the digits have good distribution properties, they conclude that this generator is predictable, and therefore not satisfactorily random, in light of their

Theorem. *From a sequence at most* $\lceil \log_b 2P^2 \rceil$ *digits, an observer can determine* P *and the next number in the sequence in polynomial time in* $\log_b P$.

Example. *Predicting the next number from the* $1/P$ *generator.*

For convenience, let's work in base $b = 10$. I'll pretend I don't know P ahead of time, so we don't know how many digits we'll need to see. For want of a better

guess, let's ask the generator for 3 digits; it responds with 407. According to **[10]**, if we write $.407 = \frac{407}{1{,}000}$ as a continued fraction and if we have enough digits, then a fraction with P as the denominator will appear as that continued fraction's first convergent whose decimal expansion matches .407 (to those three places). If that happens, then we can predict more digits correctly. We have

$$\frac{407}{1000} = 0 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{5 + \cfrac{1}{3 + \cfrac{1}{5 + \cfrac{1}{2}}}}}}$$

with convergents

$$0,\ \tfrac{1}{2} = .5,\ \tfrac{2}{5} = .4,\ \tfrac{11}{27} = .4074\ldots,\ \tfrac{35}{86} = .406\ldots,\ \tfrac{186}{457} = .4070\ldots,\ \tfrac{407}{1000} = .407$$

The first convergent that matches .407 is $11/27$ (which really can't be correct anyway because 27 isn't prime, but let's not waste time testing primality), which predicts that the next digit will be 4; however, the next digit the generator produces is 3. Repeating the process for $\frac{4{,}073}{10{,}000}$ yields the convergent $\frac{145}{356} = .40730\ldots$, which matches .4073. This predicts that the next digit will be 0, which is invalidated by the generator producing 3 next. Repeating the process for $\frac{40{,}733}{100{,}000}$ yields the convergent

$$\tfrac{200}{491} = .40733197556008146639511201629327902\ldots,$$

which matches .40733. Since the next, say, 30 digits that the generator produces match these, we accept $P = 491$. (Certainty is impossible, of course, because for any finite number of digits guessed correctly, the next one might not match.)

To present Blum, Blum, and Shub's second algorithm, the $x^2 \bmod N$ generator, we need a little notation: for an integer $N \geq 2$, let $\mathbb{Z}_N^*$ stand for the multiplicative group of invertible elements of $\mathbb{Z}/N\mathbb{Z}$ and let $\mathbb{Z}_N^{*2}$ stand for the subgroup of quadratic residues, i.e., squares, in $\mathbb{Z}_N^*$. Then, given an integer N that is the product of two distinct primes P and Q, both of which are $\equiv 3 \bmod 4$, the $x^2 \bmod N$ generator works as follows:

(1) choose an $x_0 \in \mathbb{Z}_N^{*2}$, then
(2) generate the sequence $x_{-1}, x_{-2}, x_{-3}, \ldots$ such that for every k, we have $0 < x_{-k} < N$ and

$$x_{-k-1}^2 \equiv x_{-k} \bmod N,$$

and from these
(3) obtain a sequence of *parity* bits, i.e., remainders of these mod 2.

They prove that there is an efficient algorithm for finding the parameters P, Q, and x_0 that make the period of the generated sequence long, and they prove the

Theorem.

(1) *From a knowledge of P and Q, one can efficiently compute x_{-k-1} from x_{-k} and thereby generate a sequence of bits.*

(2) *Let A be an algorithm that, given N (but not P or Q) and a portion of a sequence of bits, computes a guess for the next bit in polynomial time in $\log_2 N$ (deterministically or probabilistically). Then almost all large enough N will generate a sequence such that the fraction of the times that A guesses correctly will be only vanishingly greater than $\frac{1}{2}$.*

The proof of the second assertion assumes that N can be chosen such that the problem of factoring N into P and Q is not computable in polynomial time in $\log_2 N$; this is a widely-believed conjecture.

One interesting contrast between this (process) approach and the (sequence) approaches of Knuth and Chaitin-Kolmogorov is that the *direction* of the sequence now matters: an "observer" will find it hard to predict x_{-k-1} from x_{-k}, but not hard to predict $x_{k+1} = x_k^2 \bmod N$ from x_k. Thus, reversing a sequence negates its pseudo-randomness in this point of view but not in the other's. And, according to Yao **[12]**, the unpredictability of the $x^2 \bmod N$ generator implies that the sequence it generates passes all statistical tests, such as Knuth's, that can be computed in polynomial time.

To explain how the generator produces x_{-k-1} from x_{-k}, we need to do a little number theory. In the following, we have used x in place of x_{-k} to simplify the notation.

(1) Let $x \in \mathbb{Z}_N^*$, *where $N = PQ$ with P and Q distinct primes. Then x is congruent to a square in $\mathbb{Z}_N^*$ if and only if x is congruent to a square in both $\mathbb{Z}_P^*$ and $\mathbb{Z}_Q^*$.*
The "only if" is easy: just reduce $\bmod P$ and $\bmod Q$. To prove the "if", suppose $x \equiv y^2 \bmod P$ and $x \equiv z^2 \bmod Q$. Since P and Q are relatively prime, there are integers p and q such that $pP + qQ = 1$. Then $qQ \equiv 1 \bmod P$ and $pP \equiv 1 \bmod Q$. Therefore,

$$\pm qQy \pm pPz \equiv \begin{cases} \pm y \bmod P \\ \pm z \bmod Q. \end{cases}$$

So $(\pm qQy \pm pPz)^2 \equiv y^2 \equiv x \bmod P$ and $(\pm qQy \pm pPz)^2 \equiv z^2 \equiv x \bmod Q$, from which it follows that $(\pm qQy \pm pPz)^2 \equiv x \bmod N$, as required. It is easy to show that these four numbers are the only square roots of x in $\mathbb{Z}_N^*$.

(2) *Now assume that both P and Q are $\equiv 3 \bmod 4$. If $x \in \mathbb{Z}_N^{*2}$, let $y \equiv x^{(P+1)/4} \bmod N$ and $z \equiv x^{(Q+1)/4} \bmod N$. Then $y^2 \equiv x \bmod P$ and $z^2 \equiv x \bmod Q$.*

First we have $(y^2)^2 \equiv x^{P+1} \bmod N$, so $(y^2)^2 \equiv x^{P+1} \bmod P$. Therefore, by Fermat's little theorem **[10]**, $(y^2)^2 \equiv x^2 \bmod P$, from which it follows that $y^2 \equiv \pm x \bmod P$. Because $P \equiv 3 \bmod 4$, the quadratic reciprocity theorem **[10]** implies that -1 is not a quadratic residue mod P, and because x is a quadratic residue mod P it follows that $-x$ is not. Therefore, $y^2 \equiv x \bmod P$. The same reasoning shows that $z^2 \equiv x \bmod Q$.

(3) *With y and z as in* (2), *let $u \equiv qQy + pPz \bmod N$ and $v \equiv qQy - pPz \bmod N$ so that by* (1) *we have $(\pm u)^2 \equiv (\pm v)^2 \equiv x \bmod N$. Then exactly one of $\pm u, \pm v$ is in $\mathbb{Z}_N^{*2}$ and this is the desired x_{-k-1}.*

Let us see how $\pm u, \pm v$ must be placed in the following table:

	Square mod P	Non-square mod P
Square mod Q		
Non-square mod Q		

Quadratic reciprocity implies that u and $-u$ must be placed diagonally and so must v and $-v$. But since $v \equiv u \bmod P$ and $v \equiv -u \bmod Q$, we must place v in the same column as u but not in the same row. Therefore, exactly one of $\pm u$ and $\pm v$ occupies each box and by (1) the one that is in the upper left corner is the only one that is a quadratic residue mod N and therefore is our choice for x_{-k-1}.

Example. *Generating quadratic residues*

Let $P = 167$ and $Q = 359$, so $N = PQ = 59{,}953$. Let's start with $x_0 = 2$ and compute x_{-1}. We have $43 \times 167 + (-20) \times 359 = 1$ and $y = 8{,}337 \equiv 2^{42} \bmod 59{,}953$ and $z = 23{,}675 \equiv 2^{90} \bmod 59{,}953$. Therefore, we must determine which of the numbers

$$16{,}854 \equiv [(-20) \cdot 359 \cdot 8{,}337 + 43 \cdot 167 \cdot 23{,}675] \bmod 59{,}953$$

$$43{,}099 \equiv [20 \cdot 359 \cdot 8{,}337 - 43 \cdot 167 \cdot 23{,}675] \bmod 59{,}953$$

$$49{,}920 \equiv [(-20) \cdot 359 \cdot 8{,}337 - 43 \cdot 167 \cdot 23{,}675] \bmod 59{,}953$$

$$10{,}033 \equiv [20 \cdot 359 \cdot 8{,}337 + 43 \cdot 167 \cdot 23{,}675] \bmod 59{,}953$$

is a quadratic residue mod P and mod Q. Using quadratic reciprocity on the Jacobi symbol [10] on 16,854 with respect to $P = 167$, for example, gives

$$\left(\frac{16{,}854}{167}\right) = \left(\frac{154}{167}\right) = \left(\frac{2}{167}\right)\left(\frac{77}{167}\right) = (-1)^{\frac{167^2-1}{8}}\left(\frac{167}{77}\right)(-1)^{\frac{77-1}{2}\frac{167-1}{2}}$$

$$= \left(\frac{13}{77}\right) = \left(\frac{77}{13}\right)(-1)^{\frac{13-1}{2}\frac{77-1}{2}} = \left(\frac{12}{13}\right) = \left(\frac{-1}{13}\right) = (-1)^{\frac{13-1}{2}} = 1,$$

so 16,854 is a quadratic residue mod P. A similar calculation shows that it is also a quadratic residue mod Q and therefore also mod N, so $x_{-1} = 16{,}854$. Continuing this process, we obtain the quadratic residues

43,720 44,637 36,988 9,361 4,178 54,337 15,285 21,273 ···

which, starting with $x_{-1} = 16{,}854$, result in the sequence of parity bits beginning 001010111..., which can be shown [3] to have a period of 7298.

Turning from the mathematically unpredictable to the physically unpredictable, we describe a small electronic device that is simple and inexpensive to build; by varying one parameter, it can be made to exhibit behavior that varies from the completely predictable to the (apparently) random. In that sense, this device is essentially a coin with an adjustable level of unpredictability. The diagram of the circuit and its physical realization are shown in Figures 1 and 2.

A detailed description of its design and operation is given in [2], but we can give a general idea of how it works here, using electronic terminology that can be found in [6]. (Note that this particular design, which is not the simplest that would exhibit the same behavior, was chosen in [2] because its use of only one kind of integrated circuit simplifies the explanation of its operation.) The components represented by the triangles in the diagram are *op amps* (type 3140 with a single ended 9 volt supply) and the arrow near the top of the diagram represents a switch. The two op amps at the lower left generate a *sawtooth wave* whose frequency is determined by the voltage V at the left. The op amp at the upper middle acts as a *voltage follower* that copies this wave without disturbing its source.

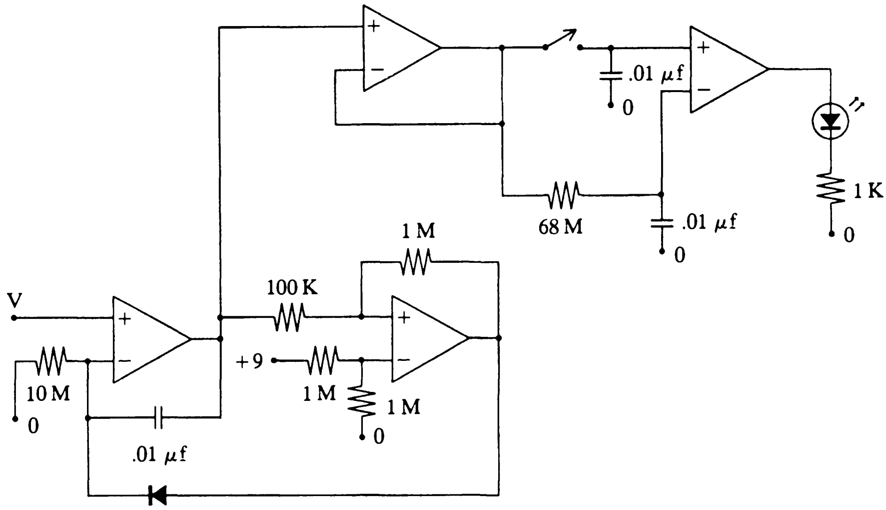


Figure 1. Diagram for an adjustable flasher.

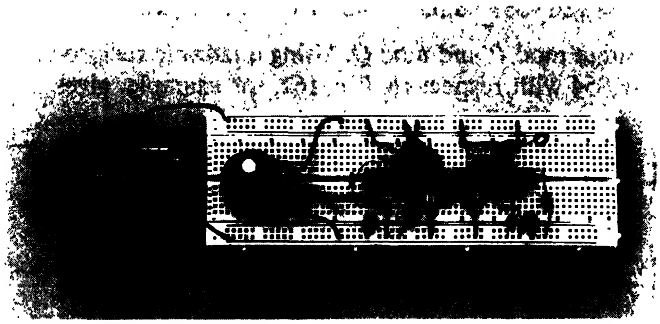


Figure 2. An adjustable flasher.

When the switch is closed, the op amp at the upper right compares the wave's instantaneous voltage with the average of its voltage, provided by the *low pass filter*, the resistor-capacitor pair connected to the "$-$" input of that op amp. This causes the LED, on the right side of the diagram, to be on half the time. When the switch is opened, the capacitor at the top of the diagram holds the voltage so the LED stays the way it was, on or off, at the moment that the switch was opened. By repeatedly closing the switch, waiting about 5 seconds to dissipate any disturbance that that might have caused, and then opening the switch, you can obtain a sequence of bits.

A low enough value of V will make the LED flash slowly enough for you to choose the moments at which you open the switch in order to get whatever sequence of on's and off's you want. A somewhat higher value of V will make that more difficult and the result will be less predictable. A still higher value of V will make the LED flash so fast that the outcome is (apparently) completely unpredictable.

It is interesting to note that the relationship between a chaotic physical system (which you can also build **[1, 2]**) and this random flasher is something like the relationship between the $1/P$ generator and the x^2 mod N generator. The behavior of a chaotic system is complicated—and therefore "looks random"—but it is

humanly predictable (in the short run) from its theory and history. In contrast, the random flasher is humanly unpredictable, but a being with greater knowledge or skill, e.g., faster reflexes, could predict the outcome. In the same vein, one could even view coin-flipping as "pseudo-random in the eyes of God": it only "looks random" to us because we don't know and can't control the initial conditions of each flip precisely enough.

The next circuit answers these objections by deriving the randomness of its display from the tiny random motion of electrons in its resistors, known as *noise* [**6**]. According to the most widely accepted interpretation of quantum mechanics [**11**], God *does* play dice with the universe on an atomic scale and the motion of each electron is completely random—it is theoretically and absolutely unpredictable independent of any possible knowledge and skill, human or divine. So while most electronic designs produce predictable results by averaging of the random motion of an incredible number of electrons—somewhere around 100,000,000,000,000,000 of them!—this circuit amplifies electron noise about 1,000,000,000 times to make the infinitesimal jiggling of smaller numbers of electrons visible as the irregular flashing of the LED. The diagram of this circuit and its physical realization are shown in Figures 3 and 4.

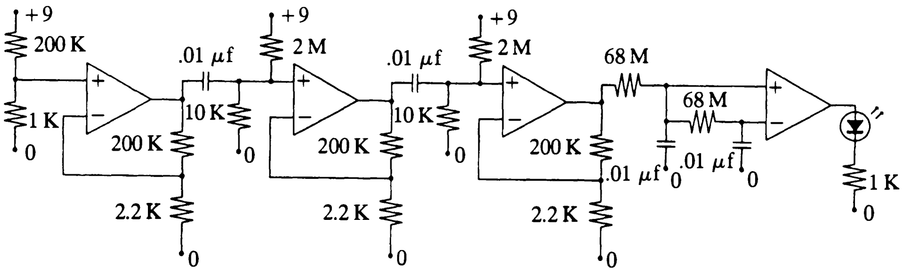


Figure 3. Diagram for the electron noise amplifier.

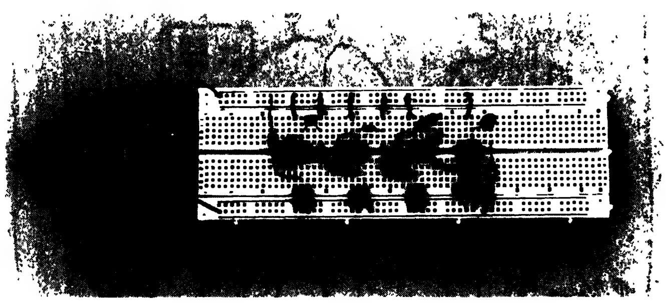


Figure 4. The electron noise amplifier.

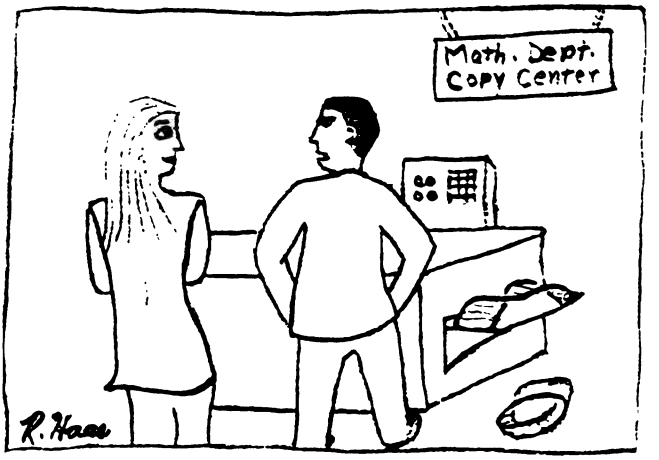
As described in detail in [**2**], the large amplification of the electron noise is accomplished by the sequence of three op amps on the left and the resistor-capacitor *high pass filters* between them, which transmit only the amplified noise, not the amplified average voltage. The op amp on the right acts as a sensitive "detector" that turns the LED on when the voltage of the amplified noise is increasing and turns it off when that voltage is decreasing.

ACKNOWLEDGMENT. It is a pleasure to acknowledge Manuel Blum's assistance with the interpretation of the $1/P$ generator.

REFERENCES

1. S. Bassein, *Order and chaos on your desk*, Amer. Math. Monthly **102**, No. 5 (1995), 409–416.
2. S. Bassein, *Mathematics Electrified!*, Grant Publishing, Berkeley (1996), 84 pp.
3. L. Blum, M. Blum, and M. Shub, *A simple unpredictable pseudo-random number generator*, SIAM J. Comput. **15**, No. 2 (1986), 364–383.
4. G. Chaitin, *Randomness and mathematical proof*, Sci. Amer. **232** (May, 1975), 47–52.
5. G. Chaitin, *A theory of program size formally identical to information theory*, JACM **22** (1972), 329–340.
6. P. Horowitz and W. Hill, *The Art of Electronics*, 2nd ed., Cambridge University Press, Cambridge (1989), 1125 pp.
7. M. Kac, *What is random*?, Amer. Scientist **71** (1983), 405–406.
8. D. E. Knuth, *The Art of Computer Programming*, 2nd ed., Addison-Wesley, Reading (1981), 689 pp.
9. A. N. Kolmogorov, *Three approaches to the quantitative definition of information*, Problems of Inform. Transmission **1** (1965), 1–7.
10. W. LeVeque, *Fundamentals of Number Theory*, Addison-Wesley, Reading (1977), 280 pp.
11. R. Omnès, *The Interpretation of Quantum Mechanics*, Princeton University Press, Princeton (1994), 550 pp.
12. A. Yao, *Theory and applications of trapdoor functions*, IEEE 23rd Symposium on Foundations of Computer Science, IEEE Computer Society Press (1982), 80–91.

Department of Mathematics and Computer Science
Mills College
Oakland, CA 94613
bassein@mills.edu



I pressed "Single-sided copy," and now it's printing everything out on Möbius strips!

Contributed by Robert Haas, Cleveland Heights, Ohio.

Mathematical Proof and the Reliability of DNA Evidence

Don Fallis

Leonard Adleman recently found the solution to a problem in graph theory using DNA technology. (See [Adleman].) However, once the solution was obtained, Adleman checked that the solution was correct by hand. As a result, DNA evidence served no justificatory function in this case. Perhaps this is fortunate since "the mathematical establishment has often expressed its displeasure with certain types of 'proof': visual, mechanical, experimental, probabilistic" [Davis 1995, p. 212]. At the risk of displeasing the mathematical establishment, I will suggest in this paper how DNA evidence might legitimately stand in for a mathematical proof.

THE DIRECTED HAMILTONIAN PATH PROBLEM. The problem in graph theory that Adleman solved is an instance of the directed Hamiltonian path problem (DHPP). A directed graph is a group of towns connected in various ways by one-way roads. One of the towns is designated the start town and another town is designated the destination town. A path is a sequence of towns such that each town in the sequence is connected by a one-way road to the next town in the sequence. A Hamiltonian path (HP) is a path from the start town to the destination town that visits every other town exactly one time. The DHPP asks if a given directed graph contains an HP and if so what the HP is.

The DHPP is an example of an NP-complete problem. A distinguishing feature of NP-complete problems is that while it is easy to check that a solution is correct, it is difficult to find a solution in the first place. In the case of the DHPP, it is easy to check whether or not any particular path is an HP, but it is difficult to find a path that is an HP.

The difficulty in solving the DHPP is not conceptual. In fact, the following simple algorithm will always find a solution. Suppose there are N towns in the graph. One by one, we generate all of the paths that visit exactly N towns. As we go, we check each path to see if it is an HP. If we find an HP, we know that there is an HP and what the HP is. If we get through all of the paths and do not find an HP, we know that there is no HP.

The difficulty in solving the DHPP is that "all known algorithms for this problem have exponential worst-case complexity, and hence there are instances of modest size for which these algorithms require an impractical amount of computer time to render a decision" [Adleman, p. 1021]. Basically, the problem is that as N increases, the number of paths that visit exactly N towns increases exponentially. Of course, this would not be a problem if we had some method for checking all of the paths at once. In essence, this is what DNA technology allowed Adleman to do.

THE DNA ALGORITHM. It is well known that DNA is a very powerful device for encoding large amounts of information. For instance, a single strand of DNA can encode all of the information necessary to generate a dinosaur. Taking advantage of this powerful device, Adleman used strands of DNA to represent the towns and roads in a graph and the paths through the graph. He then performed an algorithm consisting of the following four steps:

Step 1: Generate a whole bunch of paths through the graph.
Step 2: Remove all paths except those that begin with the start town and end with the destination town.
Step 3: Remove all paths except those that visit exactly N towns.
Step 4: Remove all paths except those that visit every town.

Any paths (i.e., strands of DNA) that remain after these four steps are performed are HPs.

A single strand of DNA is a chain of simpler molecules called bases. The four types of bases are designated A, T, G, and C. So for example, CTTGAG represents a 6-molecule-long strand of DNA. Under the right circumstances, two strands of DNA will bond together to form the familiar double helix configuration. However, since the base A will bond only with the base T and G will bond only with C, the two strands have to be the inverses of each other (where A is the inverse of T and G is the inverse of C). So, for example, two 6-molecule-long strands of DNA might bond in the following configuration:

```
CTTGAG
||||||
GAACTC
```

In the first step of the algorithm, each town in the graph is represented by a (randomly chosen) 20-molecule long strand of DNA (e.g., GCTATTCGAGCTTAAAGCTA). Thus, a path (i.e., a sequence of towns) can be represented by several town strands linked together end to end. Huge quantities of the town strands are placed in a test tube. The problem is to get them to link together in various combinations to form strands that represent paths.

This problem is solved by adding huge quantities of road strands to the test tube. Each road in the graph is also represented by a 20-molecule-long strand. The strand is designed so that it will link together the strands representing the two towns that the road connects. For instance, suppose that there is a road going from town X to town Y. The first 10 molecules in the strand representing this road are the inverses of the last 10 molecules in the town X strand; the last 10 molecules are the inverses of the first 10 molecules in the town Y strand. Thus, this road strand can bond with the second half of the town X strand and the first half of the town Y strand. As a result, this road strand can serve as "a molecular splint" [Devlin, p. 17] to hold the town X strand and the town Y strand together.

```
{       Town X       }{       Town Y       }
GCTATTCGAGCTTAAAGCTAGGCTAGGTACCAGCATGCTT
          ||||||||||||||||||||
          GAATTTCGATCCGATCCATG
          { Road  from  X  to  Y }
```

The town strands and road strands are then allowed to bond together to form double helices.[1] One strand of each double helix will consist of town strands linked together end to end and thus will represent a path through the graph. To complete the first step of the algorithm, a ligation reaction glues the town strands together so that they will not come apart even when the molecular splints (i.e., the road strands) are removed.

In the second step, a polymerase chain reaction (PCR) amplification creates multiple copies of only those paths that begin with the start town and end with the destination town. In the third step, a gel electrophoresis separation process sorts the strands of DNA by length. Only those strands that are $20 \times N$ molecules long are retained. In other words, only those strands representing paths that visit exactly N towns are retained.

In the fourth step, the test tube is heated until the double helices break apart into single strands of DNA. Next, magnetized copies of the inverse of the strand representing the first town in the graph are added to the test tube. These magnetized inverses are then allowed to bond with the single strands of DNA in the test tube. These magnetized inverses will bond only with strands representing paths that visit this first town. With the use of a magnet, these paths are retained and the rest are poured away. The test tube is again heated and the magnetized inverses are poured away (which leaves just those paths that visit this first town in the test tube). This procedure is repeated for each town in the graph. At the conclusion, the test tube contains only paths that visit each town in the graph.

Adleman used this algorithm to solve the DHPP for a relatively small graph containing only seven towns. After completing steps 1 through 4, he had a test tube that contained a number of 140-molecule-long strands of DNA, each of which represented an HP. A final PCR amplification allowed Adleman to read off the path represented by one of the strands in the test tube. Adleman then checked that the path was indeed an HP.

PROVING WITH DNA. Adleman used DNA technology to find a solution to a mathematical problem, but Adleman did not use DNA technology to justify the correctness of the solution. However, there are at least two ways in which we might try to establish the truth of a mathematical claim on the basis of DNA evidence alone. I will consider each of these two methods in turn.

First, we might try to justify the claim that there is an HP by performing only steps 1 through 4. That is, we omit reading off the path represented by one of the strands of DNA and checking that it is an HP. If the biochemical processes perform as advertised and strands of DNA remain in the test tube, then there must be an HP. In this case, the justification for the claim that there is an HP is based solely on the reliability of the DNA evidence.

I claim that, epistemically speaking, this use of DNA evidence is completely analogous to a computer proof of a mathematical claim (such as the proof of the four-color theorem). In both cases, we design a computation that will guarantee that the claim is true.[2] Next, we design a physical device (and/or process) that can

[1] This is the essence of the procedure. The actual details are slightly different. See [Adleman] for a complete description of the actual procedure.

[2] If any paths remain after steps 1 through 4, there must be an HP. Of course, if no paths remain, there may or may not be an HP.

perform this computation. Of course, the results of the computation will be reliable only if the physical device performs according to our specifications. However, we have good inductive evidence that the components of the physical device and the physical device as a whole will perform according to our specifications.

The only epistemically relevant distinction between this use of DNA evidence and a computer proof is the degree of reliability of the physical device. At least at the moment, there may be reason to believe that computer proofs are somewhat more reliable. For instance, Adleman notes that "during Step 1, the occasional ligation of incompatible edge oligonucleotides may result in the formation of molecules encoding 'pseudopaths' that do not actually occur in the graph" [Adleman, p. 1023]. If such paths were to survive the next three steps, this might lead to a false positive (i.e., strands of DNA remaining in the test tube even though there are no HPs in the graph). While Adleman claims that it is unlikely that 'pseudopaths' would survive the next three steps, he does suggest that "at the completion of a computation, it would be prudent to confirm that a putative Hamiltonian path actually occurs in the graph" [Adleman, p. 1023]. Also, since the strands of DNA representing the towns were chosen at random, it is possible (though unlikely) that strands of DNA "associated with different vertices would share long common subsequences that might result in 'unintended' binding during the ligation step (Step 1)" [Adleman, p. 1023]. Again, this might lead to a false positive.

Of course, this difference in degree of reliability is a purely contingent matter. The probability of error will undoubtedly be reduced by further advances in DNA technology. In fact, the probability of error may be no higher, even today, than you might expect with the use of a Pentium processor. (See [Markoff].) In any case, this use of DNA evidence and a computer proof are in principle epistemically equivalent. That is, there is no qualitative difference between the two methods that has any epistemic significance.

Strictly speaking, in order to reach my conclusion, I now need to argue that a computer proof (and thus DNA evidence) can legitimately stand in for a mathematical proof. However, for the following reasons, I will not do so. First, I really have nothing to add to the debate that has already been carried out on this issue. (See [Tymoczko], [Teller], and [Detlefsen].) Second, this is not a very interesting test case for this debate. It is too easy to eliminate the justificatory role of the physical process by checking the result by hand. Third and most importantly, this debate is rather anachronistic. The prevailing sentiment among mathematicians is that a computer proof is a legitimate way to establish the truth of a mathematical claim. (See, e.g., [Stewart, pp. 116–118].)

PROBABILISTIC PROVING WITH DNA. There is a second and much more interesting way in which we might try to establish the truth of a mathematical claim on the basis of DNA evidence alone. That is, we might try to justify the claim that there is no HP by performing steps 1 through 4.

There is an immediate problem with this suggestion. Even if the biological processes perform as advertised and no strands of DNA remain in the test tube, it is not necessarily the case that there is no HP. A huge number of strands of DNA representing paths are generated in the first step of the algorithm. However, there is no guarantee that all possible paths (of any given length) are generated. As a result, the biochemical process used by Adleman did not "ensure that the paths generated in Step 1 included all paths of length 7" [Devlin, p. 16]. Thus, when no

strands of DNA remain in the test tube after Step 4, it might yet be the case that an HP exists, but that a strand representing it was just not generated in Step 1.

However, even though there is no guarantee that there is no HP, there is a very high probability that there is no HP.[3] For each town and each road, Adleman put 50 picomoles of the associated strand of DNA in the test tube in Step 1. This amounted to putting "approximately 3×10^{13} copies" [Adleman, p. 1022] of each town strand and each road strand in the test tube. According to Adleman, "the quantity used should be just sufficient to insure that during the ligation step (Step 1) a molecule encoding a Hamiltonian path will be formed with high probability if such a path exists in the graph" [Adleman, p. 1023]. As a result, if no strand representing an HP is generated in Step 1, then the probability that there is no HP in the graph is very high. In fact, "for all practical purposes this result could be taken as certainty, since most 'definite' conclusions in everyday life are based on far lower probabilities" [Devlin, p. 16].

Nevertheless, the prevailing sentiment among mathematicians is that a probabilistic verification is not a legitimate substitute for a mathematical proof. For instance, the computer scientist David Harel claims that "as long as we use probabilistic algorithms only for petty, down-to-earth matters such as wealth, health, and survival, we can easily make do with very-likely-to-be-correct answers to our questions. The same, it seems, cannot be said for our quest for absolute mathematical truths" [Harel, p. 332]. I claim that mathematicians do not have good grounds for this rejection of probabilistic verifications in general and of probabilistic DNA proofs (PDP) in particular. More precisely, I claim that there is no epistemically important difference between the methods acceptable to mathematicians and a PDP.[4]

For instance, the difference is not that methods acceptable to mathematicians are infallible and that a PDP is not. A PDP is clearly not infallible, but neither are the methods acceptable to mathematicians. For instance, according to the mathematician Philip Davis, "a mathematical error of international significance may occur every twenty years or so" [Davis 1972, p. 262] and those are just the errors that get written up in the newspapers. In fact, if steps 1 through 4 are repeated a number of times and no strands ever remain in the test tube, a PDP will arguably provide more reliable evidence than many of the methods acceptable to mathematicians (such as writing down a 10,000 page proof).

The difference is not that methods acceptable to mathematicians involve the construction of a "complete proof" that leaves "no gaps, no loopholes, no uncertainty whatever" [Polya, p. 215] and that a PDP does not. A PDP does not involve the construction of a rigorous mathematical proof (that there are no HPs in a graph), but neither do some methods acceptable to mathematicians. For instance, in 1931, Gödel established (to the satisfaction of the mathematical community) that no consistent axiomatization of arithmetic can prove its own consistency. (See [Dawson].) However, it was another 8 years before anyone wrote down a rigorous mathematical proof of the second incompleteness theorem.

[3]In fact, this probability may be higher than the probability that there is an HP when using DNA evidence alone. For instance, in this case, false positives are not a problem.

[4]The support I will offer for this claim applies equally well to many other probabilistic techniques, such as probabilistic primality tests (e.g., see [Rabin] or [Solovay]).

The difference is not that methods acceptable to mathematicians provide good evidence that we could construct a rigorous mathematical proof and that a PDP does not. As we saw above, a PDP provides good evidence that there are no HPs in a graph. If there are no HPs in a graph, then it is a trivial (but perhaps tedious) matter to write down a proof of this fact. For instance, we could perform the simple algorithm (which generates all of the paths that visit exactly N towns) described above. Thus, a PDP does provide good evidence that we could construct a rigorous mathematical proof.

The difference is not that methods acceptable to mathematicians are a priori justifications and that a PDP is not. There is a sense in which a PDP is not an a priori justification. That is, it makes an appeal to empirical data (e.g., that strands of DNA bond together in particular ways). However, the proof of the four-color theorem is not a priori in this sense either. For instance, the proof of the four-color theorem appeals to the empirical fact that transistors behave in a particular way.

Of course, there is a sense in which the proof of the four-color theorem is an a priori justification. It need not appeal to any particular empirical data and in principle need not appeal to empirical data at all. For instance, the relevant computations could be performed by a device other than a digital computer and in principle could be performed in the mathematician's mind. However, a PDP is a priori in this sense as well. Steps 1 through 4 could be performed by a computer program running on a digital computer (instead of by strands of DNA in a test tube) and in principle could be carried out in the mathematician's mind.

Of course, the preceding paragraphs do not provide an exhaustive demonstration that there is no epistemically important difference between the methods acceptable to mathematicians and a PDP. (See [Fallis] for a more exhaustive demonstration.) However, while there may be "a qualitative difference between probabilistic verification and mathematical proof that is important to mathematicians" [Pomerance, p. 142], it is not immediately clear what this difference is and whether or not it has any epistemic significance.[5] Until and unless an epistemically relevant difference is clearly identified, a PDP is arguably as legitimate a method of establishing the truth of a mathematical claim as the proof of the four-color theorem or arguments such as Gödel's (which if there is any truth in advertising should end with "Quod erat sketchandum" [Steel, p. 83] instead of QED).

REFERENCES

Adleman, Leonard M., "Molecular Computation of Solutions to Combinatorial Problems," *Science*, v. 266, 1994, pp. 1021–1024.

Davis, Philip J., "Fidelity in Mathematical Discourse: Is One and One Really Two?", *The American Mathematical Monthly*, v. 79, no. 3, 1972, pp. 252–263.

Davis, Philip J., "The Rise, Fall, and Possible Transfiguration of Triangle Geometry: A Mini-history", *The American Mathematical Monthly*, v. 102, no. 3, 1995, pp. 204–214.

Dawson, John W., "The Reception of Gödel's Incompleteness Theorems", *PSA 1984*, v. 2, 1985, pp. 253–271.

Detlefsen, Michael, and Mark Lukar, "The Four-Color Theorem and Mathematical Proof", *Journal of Philosophy*, v. 76, 1980, pp. 803–820.

Devlin, Keith, "Test Tube Computing with DNA", *Math Horizons*, April 1995, pp. 14–21.

Fallis, Don T., "A Defense of a Probabilistic Method of Establishing Mathematical Truths", Ph.D. thesis, University of California, Irvine, 1995.

[5]By the way, I take it that merely pointing out that a PDP is probabilistic just begs the question.

Harel, David, *Algorithmics: The Spirit of Computing*, Addison-Wesley, 1987.
Markoff, John, "Flaw undermines accuracy of Pentium chips", *New York Times*, Nov. 24, 1994, p. D1.
Polya, G., *How to Solve It*, 2nd edition, Princeton University Press, 1957.
Pomerance, Carl, "The Search for Prime Numbers", *Scientific American*, Dec. 1982, pp. 136–147.
Rabin, M. O., "Probabilistic Algorithm for Testing Primality", *Journal of Number Theory*, v. 12, 1980, pp. 128–138.
Solovay, R., and V. Strassen, "A Fast Monte-Carlo Test for Primality", *SIAM Journal on Computing*, v. 6, no. 1, 1977, pp. 84–85.
Steel, John R., "$\mathrm{HOD}^{L(R)}$ is a core model below O", *The Bulletin of Symbolic Logic*, v. 1, no. 1, 1995, pp. 75–84.
Stewart, Ian, *The Problems of Mathematics*, Oxford University Press, 1987.
Teller, Paul, "Computer Proof", *Journal of Philosophy*, v. 76, 1980, pp. 797–803.
Tymoczko, Thomas, "The Four-Color Problem and its Philosophical Significance", *Journal of Philosophy*, v. 76, no. 2, 1979, pp. 57–83.

Department of Philosophy
University of California, Irvine
Irvine, CA 92717
ghumdfp@benfranklin.hnet.uci.edu

Logic does not say anything about the world but has to do only with the way in which I talk about the world. . . .

If someone does not want to accept logical inference, it is not that he has a different opinion from mine about the behaviour of objects, but that he is refusing to talk about objects according to the same rules as I; it is not that I cannot convince him, but that I must refuse to go on talking with him, just as I shall refuse to go on playing tarot with a partner who insists on taking my fool with the moon.

Language correlates combinations of symbols with states of affairs in the world, and the way it correlates them is not one-to-one (which would be quite pointless) but many-to-one; and logic gives the rules about the way in which one combination of symbols can be transformed into another one which designates the same state of affairs; this is what is called the tautological character of logic.

It seems hardly credible at first sight that the whole of mathematics with its hard-earned theorems and its frequently surprising results could be dissolved into tautologies. But this argument overlooks just a minor detail, namely the circumstance that we are not omniscient.

An omniscient subject needs no logic, and contrary to Plato we can say: God never does mathematics.

Hans Hahn, as quoted by Karl Sigmund in
A Philosopher's Mathematician: Hans Hahn and the Vienna
Circle, *The Mathematical Intelligencer* 17 (4) (1995), p.26.

NOTES

Edited by: **John Duncan**

Equidissections of Trapezoids

Charles H. Jepsen

Suppose a plane polygon is dissected into triangles of equal areas. What numbers of triangles are possible? Given a polygon K, a dissection of K into m triangles of equal areas is called an *m-equidissection*. The *spectrum* of a polygon K, denoted $S(K)$, is the set of integers m such that K has an m-equidissection. If $S(K)$ consists of all multiples of a single integer m, we say that $S(K)$ is *principal* and write $S(K) = \langle m \rangle$. Monsky **[2]** showed that if K is a square, then $S(K) = \langle 2 \rangle$. Kasimatis **[1]** showed that if K is a regular n-gon with $n \geq 5$, then $S(K) = \langle n \rangle$. A discussion of this problem for other polygons is contained in Chapter 5 of **[4]**.

What can we say about the spectrum of a trapezoid? We may confine our attention to trapezoids with vertices $(0,0)$, $(1,0)$, $(0,1)$, $(a,1)$, $a > 0$, since any trapezoid is affinely equivalent to such a trapezoid. Denote this trapezoid by $T(a)$.

The following are shown in **[4]** (pp. 121–122):

If a is rational, say $a = r/s$ where r and s are relatively prime positive integers, then $S(T(a)) = \langle r + s \rangle$.

If a is transcendental, then $S(T(a))$ is the empty set.

Two questions raised on p. 126 of **[4]** are:

(1) What is the spectrum of the trapezoid $T(\frac{1}{3}(6 + \sqrt{21}))$?
(It is shown that $\langle 20 \rangle \subseteq S(T(\frac{1}{3}(6 + \sqrt{21}))) \subseteq \langle 10 \rangle$.)

(2) Is the spectrum of every trapezoid principal?

We answer these questions as follows:

(1) $10 \in S(T(\frac{1}{3}(6 + \sqrt{21})))$ so $S(T(\frac{1}{3}(6 + \sqrt{21}))) = \langle 10 \rangle$.

(2) There are infinitely many values of a such that $S(T(a))$ is *not* principal.

Our main tool is the following (Theorem 4, p. 123 in **[4]**):

Theorem. *Let t_1, t_2, t_3 be positive integers such that $t_2^2 - 4t_1t_3$ is positive and is not the square of an integer. Let a be one of the roots of the equation $t_3x^2 - t_2x + t_1 = 0$. Then $T(a)$ has an equidissection into $t_1 + t_2 + t_3$ triangles.*

To prove (1), let $a = \frac{1}{3}(6 + \sqrt{21})$. Then $t_1 = 5$, $t_2 = 12$, $t_3 = 3$, and the trapezoid $T(a)$ has a 20-equidissection. We find a dissection into 10 triangles, each of area $\frac{1}{60}(9 + \sqrt{21})$. The line $x = \frac{1}{30}(9 + \sqrt{21})$ partitions $T(a)$ into a rectangle and a trapezoid T. The rectangle forms two right triangles of area $\frac{1}{60}(9 + \sqrt{21})$. The ratio of the lengths of the sides of T is $\frac{1}{7}(21 + 4\sqrt{21})$. Hence T is affinely equivalent to

the trapezoid $T(\frac{1}{7}(21 + 4\sqrt{21}))$. We dissect $T(\frac{1}{7}(21 + 4\sqrt{21}))$ into 8 triangles as follows.

The line joining $(1, 0)$ to a point $(b, 1)$ partitions $T(\frac{1}{7}(21 + 4\sqrt{21}))$ into a triangle S of area A_1 and a trapezoid $T(b)$ of area A_2. Choose b so that $A_1/A_2 = 1/7$; then b satisfies the equation $\frac{1}{2}(\frac{1}{7}(21 + 4\sqrt{21}) - b)/\frac{1}{2}(1 + b) = \frac{1}{7}$. Upon simplifying, we see that b is a root of the equation $x^2 - 5x + 1 = 0$. It follows from the theorem that $T(b)$ can be dissected into $1 + 5 + 1 = 7$ triangles, each of area $A_2/7 = A_1$. These seven triangles, together with S, form a dissection of $T(\frac{1}{7}(21 + 4\sqrt{21}))$ into 8 triangles. Hence $T(\frac{1}{3}(6 + \sqrt{21}))$ has an equidissection into 10 triangles.

We must look elsewhere for values of a such that $S(T(a))$ is not principal. The following theorem gives an infinite collection of such values of a.

Theorem. *Let $a = r + \sqrt{r^2 - 3}$ where r is an even integer, $r \geq 4$. Then the spectrum of the trapezoid $T(a)$ is not principal.*

Proof: We show that $2(r + 2)$ and $(r + 1)(r + 2)$ lie in $S(T(a))$ but $r + 2$ is not in $S(T(a))$. Since a is a root of $x^2 - 2rx + 3 = 0$, by the preceding theorem, $T(a)$ can be dissected into $3 + 2r + 1 = 2(r + 2)$ triangles. Following our previous argument, use the point $(b, 1)$ to partition $T(a)$ into a triangle S of area A_1 and a trapezoid $T(b)$ of area A_2 so that $A_1/A_2 = 1/(r + 1)$. Then b satisfies the equation $\frac{1}{2}(a - b)/\frac{1}{2}(1 + b) = 1/(r + 1)$. Simplifying, we find that b is a root of the equation $(r/2 + 1)x^2 - (r^2 + r - 1)x + (r/2 + 1) = 0$. Hence $T(b)$ can be dissected into $(r/2 + 1) + (r^2 + r - 1) + (r/2 + 1) = (r + 1)^2$ triangles, each of area $A_2/(r + 1)^2 = A_1/(r + 1)$. Since S can be dissected into $r + 1$ triangles of area $A_1/(r + 1)$, we have an equidissection of $T(a)$ into $(r + 1)^2 + (r + 1) = (r + 1)(r + 2)$ triangles. Because $r + 1$ is odd, the greatest common divisor of $2(r + 2)$ and $(r + 1)(r + 2)$ is $r + 2$. Thus $S(T(a))$ is principal only if it contains $r + 2$. Suppose $T(a)$ is dissected into $r + 2$ triangles. Then the area of each triangle is

$$\begin{aligned}\frac{1}{2(r+2)}(1 + a) &= \frac{1}{2(r+2)}\left(1 + r + \sqrt{r^2 - 3}\right)\\ &> \frac{1}{2(r+2)}(1 + r + (r - 1)) = \frac{r}{r+2} > \frac{1}{2},\end{aligned}$$

since $r > 2$. But this dissection must contain a triangle whose base lies along the base of the trapezoid; the area of such a triangle is at most $\frac{1}{2}$. This contradiction shows that $T(a)$ cannot be dissected into $r + 2$ triangles. Consequently, $S(T(a))$ is not principal. ∎

Let $a = p + q\sqrt{d}$ where p and q are positive rational numbers, d is a square-free positive integer. The following questions about $S(T(a))$ remain open:

(1) When is $S(T(a))$ the empty set?

$$\left(\text{Conjecture: } S\left(T\left(p + q\sqrt{d}\right)\right) = \varnothing \text{ iff } p - q\sqrt{d} < 0.\right)$$

(2) When is $S(T(a))$ principal?

When $S(T(a))$ is not principal, this spectrum is in general difficult to determine explicitly. However, building on the above ideas, Monsky shows (see the following note [3]) that $S(T(15 + 8\sqrt{3}))$ consists of all multiples of 8 other than 8, 16, and 24.

REFERENCES

1. E. A. Kasimatis, Dissections of regular polygons into triangles of equal areas, *Discrete and Computational Geometry* **4** (1989), 375–381.
2. P. Monsky, On dividing a square into triangles, *Amer. Math. Monthly* **77** (1970), 161–164.
3. P. Monsky, Calculating a trapezoidal spectrum, *Amer. Math. Monthly* **103** (1996), 500–501
4. S. K. Stein and S. Szabó, *Algebra and Tiling: Homomorphisms in the Service of Geometry*, Mathematical Association of America, Washington, DC, 1994.

Department of Mathematics
Grinnell College
Grinnell, IA 50112-0806
jepsen@math.grin.edu

Calculating a Trapezoidal Spectrum

Paul Monsky

In [**1**], whose language and notation I adopt, Charles Jepsen has constructed trapezoids, $T(a)$, with non-principal spectrum. It's far from clear what such a spectrum can look like, but I'll work out one example, showing:

Theorem. *Let r be an integer ≥ 8. Set $a = (2r - 1) + r\sqrt{3} = ((r - 2) + \sqrt{3})/(2 - \sqrt{3})$. Then $S(T(a))$ consists of all multiples of r other than r, $2r$, and $3r$.*

Proof: Suppose $T(a)$ has an n-equidissection. Let p be a prime and let φ_p be a valuation on the reals extending the p-adic valuation on Q. By [**2**] (page 118), $\varphi_p(n) \geq \varphi_p(2 \text{ area } T(a)) = \varphi_p(r(2 + \sqrt{3})) = \varphi_p(r)$, so r divides n. Furthermore, one of the triangles in the equidissection has its base on the x-axis. The area of such a triangle is $\leq 1/2$; it follows that $n \geq r(2 + \sqrt{3}) > 3r$. It remains to show that $T(a)$ has an mr-equidissection for each $m \geq 4$. ∎

Let $u = [m\sqrt{3}]$. Since $m \geq 4$, $2m - u - 1 \geq 0$. Also $u^2 + u - 3m^2 < (m\sqrt{3} + 1)(m\sqrt{3} + 2) - 3m^2 = 3\sqrt{3}m + 2$. Since $r \geq 8$, this is no more than $m(r - 2)$, and $m(r - 2) - (u^2 + u - 3m^2) > 0$. Note also that $(u + m\sqrt{3})(u + 1 - m\sqrt{3}) = u^2 + u - 3m^2 + m\sqrt{3}$.

Now let T be a trapezoid of height 2 with bases equal to $u^2 + u - 3m^2 + m\sqrt{3}$ and $u + 1 - m\sqrt{3}$; the last sentence shows that T is affine equivalent to $T(u + m\sqrt{3})$. Since $u + m\sqrt{3}$ is a root of $x^2 - 2ux + u^2 - 3m^2$, the result of Stein quoted in [**1**] tells us that T may be dissected into $u^2 + 2u + 1 - 3m^2$ equal triangles. Each of these triangles has area 1. Let T^* be a trapezoid formed from T by adding a triangle of height 2 along each slant side of T. The bases of these two triangles are chosen to be $m(r - 2) - (u^2 + u - 3m^2)$ and $2m - u - 1$, integers

that we have shown to be positive. Then T^* has height 2 and bases equal to $m(r-2+\sqrt{3})$ and $m(2-\sqrt{3})$. T^* has area mr, and evidently can be dissected into triangles of area 1. Since $(r-2+\sqrt{3})/(2-\sqrt{3})=a$, T^* is affine equivalent to $T(a)$, and $T(a)$ admits an mr-equidissection.

REFERENCES

1. C. H. Jepsen, Equidissections of Trapezoids, *Amer. Math Monthly* **103** (1996), 498–500.
2. S. K. Stein and S. Szabó, Algebra and Tiling: Homomorphisms in the Service of Geometry, Mathematical Association of America, Washington, DC, 1994.

Department of Mathematics
Brandeis University
Waltham, MA 02254-9110
monsky@binah.cc.brandeis.edu

Reverend Charles to the aid of Major Percy and Fields Medalist Enrico

Doron Zeilberger

Voltaire said that Archimedes had more imagination than Homer. Unfortunately, most of mathematicians' creativity can be appreciated only by mathematicians themselves. Sometimes, however, mathematicians employ their imagination to do non-mathematical activity. Notable examples are Multi-Millionaire Richard Garfield, the Reverend Charles Dodgson, and Major Percy MacMahon, who respectively developed: 'Magic: The Gathering' (*the* game of our decade), Alice, and an early version of Instant Insanity.

This is not to say that their imagination did not also help mathematics proper. In this *quickie*, I observe how Dodgson's rule for evaluating determinants **[D]** (for any $n \times n$ matrix A, let $A_r(k,l)$ be the $r \times r$ submatrix whose upper leftmost corner is the entry $a_{k,l}$),

$$\det A = \frac{\det A_{n-1}(1,1)\det A_{n-1}(2,2) - \det A_{n-1}(1,2)\det A_{n-1}(2,1)}{\det A_{n-2}(2,2)},$$

immediately implies MacMahon's determinant evaluation **[M]**:

$$\det\left[\binom{a+i}{b+j}_{1\le i,j\le n}\right] = \frac{(a+n)!!(n-1)!!(a-b-1)!!(b)!!}{(a)!!(a-b+n-1)!!(b+n)!!},$$

where, $n!! := 1!2!3!\cdots n!$, and, of course, $n! := 1\cdot 2\cdots n$.

Indeed, let the left and right sides be $L_n(a,b)$ and $R_n(a,b)$, respectively. Dodgson's rule immediately implies that the recurrence

$$X_n(a,b) = \frac{X_{n-1}(a,b)X_{n-1}(a+1,b+1) - X_{n-1}(a+1,b)X_{n-1}(a,b+1)}{X_{n-2}(a+1,b+1)}$$

holds with $X = L$. Since $L_n(a, b) = R_n(a, b)$ for $n = 1, 2$ (check!), and the recurrence also holds with $X = R$ (check![1]), it follows by induction that $L_n(a, b) = R_n(a, b)$ for *all* n. ■

The special case $a = 2n + 1$, and $b = n$, reduces to a special case of a conjecture of Enrico Bombieri[2], David Hunt, and Alf van der Poorten ([**BHP**]). This special case was done in [**BHP**] using a different method, but I believe that Dodgson's method can be extended to prove their full conjecture.

REFERENCES

[BHP] E. Bombieri, D. C. Hunt, and A. J. van der Poorten, Determinants in the study of Thue's method and curves with prescribed singularities, *J. Experimental Mathematics*, to appear.

[D] C. L. Dodgson, Condensation of Determinants, *Proceedings of the Royal Society of London* **15** (1866), 150–155.

[M] P. A. MacMahon, *Combinatory Analysis*, vol. 2, Cambridge University Press, 1918. [reprinted by Chelsea, 1984].

Department of Mathematics
Temple University
Philadelphia, PA 19122
zeilberg@math.temple.edu

On Accumulation Points of Ratio Sets of Positive Integers

József Bukor and János T. Tóth

It is the purpose of this note to answer the first part of the third open problem in [**1**] and make some comments concerning [**4**]. Denote by $\mathbb{R}$ ($\mathbb{R}^+$) the real (positive real) numbers and by $\mathbb{Q}^+$ and $\mathbb{N}$ the positive rationals and the positive integers, respectively. The ratio set of $A \subset \mathbb{N}$ is denoted by $R(A) = \{\frac{a}{b}; a, b \in A\}$ (see [**2**], [**3**], [**5**] for this notion; in [**1**] and [**4**] the symbol $F(A)$ is used). Let X^d stand for the set of all accumulation points of $X \subset \mathbb{R}^+$. The integer part of $x \in \mathbb{R}$ is denoted by $\lfloor x \rfloor$.

In [**1**] it was asked: For which sets $B \subset \mathbb{R}$ does there exist a set $A \subset \mathbb{N}$ such that $R(A)^d = B$?

It is evident that $B \neq \emptyset$ provided A is infinite. On the other hand, $\{0, +\infty\} \subset R(A)^d$ for any infinite $A \subset \mathbb{N}$. Further, if some positive $t \in R(A)^d$, then $\frac{1}{t} \in R(A)^d$, since $\frac{a}{b} \in R(A)$ always implies that $\frac{b}{a} \in R(A)$. Notice also that the accumulation points of any linear set constitute a closed set in $\mathbb{R}$. Consequently, the nonempty

[1] Divide both sides by the left, then use $r!!/(r-1)!! = r!$ whenever possible, and then $r!/(r-1)! = r$ whenever possible, reducing it to a completely routine polynomial identity.

[2] Who applies his imagination not only to mathematics, but also to art: he is an accomplished painter.

set B we are looking for in the above open problem must be a closed subset of $[0, +\infty] = \mathbb{R}^+ \cup \{0, +\infty\}$, it must contain 0 and $+\infty$, and if $b \in B$ ($b \in \mathbb{R}^+$) then $\frac{1}{b} \in B$. As the following theorem shows, these conditions are also sufficient for B:

Theorem. *Let $\emptyset \neq B \subset [0, +\infty]$. The following are equivalent:*

(i) *There exists an $A \subset \mathbb{N}$ such that $R(A)^d = B$;*
(ii) *$B \cap \mathbb{R}$ is closed in $\mathbb{R}$, $\{0, +\infty\} \subset B$, and $b \in B \cap \mathbb{R}^+$ implies $\frac{1}{b} \in B$.*

Proof: It suffices to prove only (ii) $\Rightarrow$ (i). Let $\emptyset \neq B \subset [0, +\infty]$ and suppose B satisfies (ii). Let $\mathcal{S}$ stand for the system of intervals $(1 + (i-1)/n, 1 + (i+1)/n)$ where $n \in \mathbb{N}$ and $i = 1, 2, \dots, n^2$. The length of intervals tends to zero with increasing n and every real number greater than 1 can be covered with infinitely many elements of $\mathcal{S}$. Denote by $((c_k - \delta_k, c_k + \delta_k))_{k=1}^{\infty}$ the sequence of those intervals from $\mathcal{S}$ that meet B (i.e., that contain at least one element from B).

Define the set $A = \{a_0 < a_1 < a_2 < \cdots\} \subset \mathbb{N}$ as follows: Let $a_0 = 1$, further

$$a_{2n+1} = 2^{a_{2n}} \quad \text{and} \quad a_{2n+2} = \lfloor a_{2n+1} \cdot c_n \rfloor \text{ for all } n = 0, 1, 2 \dots.$$

We will show that $R(A)^d = B$.

(1) $B \subset R(A)^d$: Let $t \in B$ be a positive real number. We may suppose that $t > 1$. Let $((c_{n_k} - \delta_{n_k}, c_{n_k} + \delta_{n_k}))_{k=1}^{\infty}$ be a sequence of intervals containing t. Then $\lim_{k \to \infty} c_{n_k} = t$ since $\lim_{k \to \infty} \delta_{n_k} = 0$. Accordingly, the sequence

$$\frac{a_{2n_k+2}}{a_{2n_k+1}} = \frac{\lfloor a_{2n_k+1} c_{n_k} \rfloor}{a_{2n_k+1}} \quad (k \in \mathbb{N})$$

converges to t; thus, $t \in R(A)^d$.

(2) $R(A)^d \subset B$: Let $z \in R(A)^d$ and $z > 1$. Then there exist sequences $(m_k)_{k=1}^{\infty}$ and $(l_k)_{k=1}^{\infty}$ of positive integers such that

$$\lim_{k \to \infty} \frac{a_{m_k}}{a_{l_k}} = z. \tag{1}$$

Observe that $\lim(a_{2n+1}/a_{2n}) = +\infty = \lim(a_{2n+3}/a_{2n+1})$, so (1) can hold only if there is a sequence $(n_k)_{k=1}^{\infty}$ of positive integers with

$$\lim_{k \to \infty} \frac{a_{2n_k+2}}{a_{2n_k+1}} = z. \tag{2}$$

Since $a_{2n_k+2} = \lfloor a_{2n_k+1} c_{n_k} \rfloor$, it follows from (2) that $\lim_{k \to \infty} c_{n_k} = z$, hence $\lim_{k \to \infty}(c_{n_k} - \delta_{n_k}) = \lim_{k \to \infty}(c_{n_k} + \delta_{n_k}) = z$. Further, every interval $(c_{n_k} - \delta_{n_k}, c_{n_k} + \delta_{n_k})$ contains some $t_k \in B$; therefore, $\lim_{k \to \infty} t_k = z$. Finally, the closedness of $B \cap \mathbb{R}$ in $\mathbb{R}$ ensures that $z \in B$. ∎

In the remaining part we will make some comments concerning Theorem 1 in [4], which states that if there exists a strictly increasing sequence $(a_n)_{n=1}^{\infty}$ in A with $\lim_{n \to \infty}(a_{n-1}/a_n) = 1$, then $R(A)$ is dense in $\mathbb{R}^+$. This theorem represents a direct proof, in relation to the problem in question, of a more general result due to Narkiewicz and Šalát ([2], Proposition 1 and 2). On the other hand, it may be of interest that in view of a theorem of Zsilinszky in [5] if a set $A = \{a_1 < a_2 < \cdots\} \subset \mathbb{R}^+$ is such that $\liminf_{n \to \infty}(a_{n+1}/a_n) = c$, where $c \in (1, +\infty]$, then $R(A)$ is not dense in $\mathbb{R}^+$; in fact, $R(A)^d \cap (1/c, c) = \emptyset$.

Our final observation concerns the first open problem in [1] and Theorem 1 in [4]. It seems improbable to characterize sets $A \subset \mathbb{N}$ for which $R(A)$ is dense in $\mathbb{R}^+$ by conditions involving limits as in [4] Theorem 1.

Indeed, as the subsequent example demonstrates, if we are given an infinite set $A \subset \mathbb{N}$, it is sufficient to add (in comparison with A) only a "few" elements to A and the resulting set will already be dense in $\mathbb{R}^+$. More precisely, if the elements of A do not satisfy some equality involving limits, then the resulting set will not satisfy it either; however, its ratio set will be dense in $\mathbb{R}^+$.

Example. Let $a_1 < a_2 < \cdots$ be an arbitrary infinite sequence of positive integers. Choose an arbitrary subsequence $(a_{n_k})_{k=1}^{\infty}$ of $(a_n)_{n=1}^{\infty}$. Denote by $(r_n)_{n=1}^{\infty}$ the sequence containing every rational number greater than 1 infinitely many times. Let $b_k = \lfloor a_{n_k} \cdot r_k \rfloor$ for $k = 1, 2 \dots$. We will show that the ratio set of $C = \{a_k; k \in \mathbb{N}\} \cup \{b_k; k \in \mathbb{N}\}$ is dense in $\mathbb{R}^+$. It suffices to show that $\mathbb{Q}^+ \subset R(C)^d$.

Let $s \in \mathbb{Q}^+$. Without loss of generality, we may assume that $s > 1$. In view of the definition of $(r_n)_{n=1}^{\infty}$, we get a sequence $(m_l)_{l=1}^{\infty}$ of positive integers such that $r_{m_l} = s$ for all $l \in \mathbb{N}$. Then

$$\frac{b_{m_l}}{a_{n_{m_l}}} = \frac{\lfloor a_{n_{m_l}} \cdot s \rfloor}{a_{n_{m_l}}} \to s \text{ as } l \to \infty;$$

thus, $s \in R(C)^d$.

REFERENCES

1. D. Hobby, D. M. Silberger, Quotients of primes, *Amer. Math. Monthly* **100** (1993), 50–52.
2. W. Narkiewicz, T. Šalát, A theorem of H. Steinhaus and (R)-dense sets of positive integers, *Czechoslov. Math. J.* **34** (109) (1984), 355–361.
3. T. Šalát, On ratio sets of natural numbers, *Acta Arith.* **15** (1969), 273–278.
4. P. Starni, Answers to two questions concerning quotients of primes, *Amer. Math. Monthly* **102** (1995), 347–349.
5. J. T. Tóth, L. Zsilinszky, On density of ratio sets of powers of primes, *Nieuw Archief voor Wiskunde* **13** (1995), 205–208.

Department of Mathematics
University of Education
Farská 19, 949 74 Nitra, Slovakia
e-mail: bukor@unitra.sk, toth@unitra.sk

Answer to Picture Puzzle

(p. 482)

Kennan T. Smith and William F. Donoghue in 1969

UNSOLVED PROBLEMS

Edited by: **Richard Guy & Richard Nowakowski**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics & Statistics, The University of Calgary, Alberta, Canada T2N 1N4.

A Polygon Problem

G. C. Shephard

Given a simple polygon, it is clearly of interest to determine the number of points of intersection of its diagonals. Here we consider the opposite problem: how many sets of diagonals of a given polygon are disjoint, that is, have pairwise empty intersections? We pose two problems ((i) and (ii) below) concerning identities and inequalities between these numbers. But first we need some definitions.

A polygon with n vertices is said to be *simple* if the intersection of every two distinct edges is either empty, or, if the edges are contiguous, is an end-point of each (a vertex of $P(n)$). Such a polygon has a well-defined interior and exterior.

A *chord* is a relatively open line-segment whose end-points are non-consecutive vertices of $P(n)$. Two chords are disjoint if their intersection is empty; in particular, distinct chords are disjoint even if they have an end-point in common. A chord is called a *diagonal* if it lies in the interior of $P(n)$ and an *epigonal* if it lies in the exterior of $P(n)$. Let d_1 be the number of diagonals, d_2 be the number of disjoint pairs of diagonals, and, in general, d_i the number of sets of i diagonals of $P(n)$ which are pairwise disjoint. The numbers e_i are defined in a similar manner for epigonals. It is not hard to show that $d_i = 0$ and $e_i = 0$ if $i \geq n-2$, and that if the diagonals of $P(n)$ intersect in distinct points, then the number of such points is $\frac{1}{2}(d_1(d_1-1)-d_2)$. This is illustrated in the diagram for various hexagons ($n = 6$).

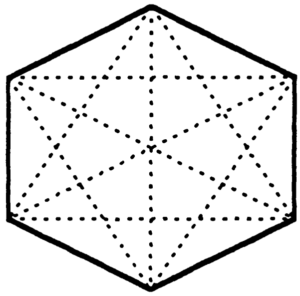
If $P(n)$ is strictly convex (that is, is convex and no three consecutive vertices are collinear), Lee **[1]** has shown that there exists an $(n-3)$-dimensional polytope A_n, called the **associahedron**, whose d_1 vertices correspond to the diagonals of $P(n)$ and whose d_{n-3} facets ($(n-4)$-dimensional faces) correspond to the triangulations of $P(n)$. Applying Euler's Theorem to A_n it is immediate that

$$d_1 - d_2 + d_3 - \cdots + (-1)^n d_{n-3} = 1 + (-1)^n.$$

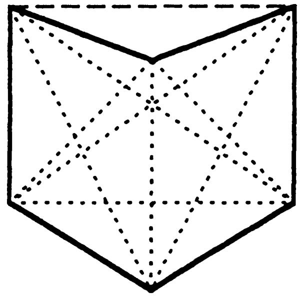
It would be interesting to find a direct combinatorial proof of this identity.

(i) *Show that each of the equalities*

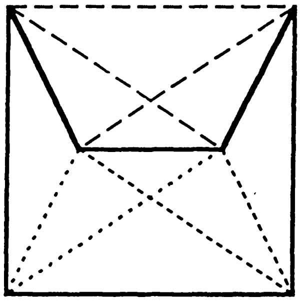
$$\sum_{i=1}^{n-3} (-1)^{i-1} d_i = 1 \qquad \textit{and} \qquad \sum_{i=1}^{n-3} (-1)^{i-1} e_i = 1$$



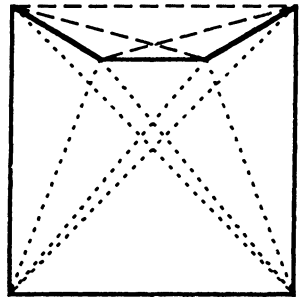
$d_1 = 9$
$d_2 = 21$
$d_3 = 14$
$e_1 = 0$
$e_2 = 0$
$e_3 = 0$



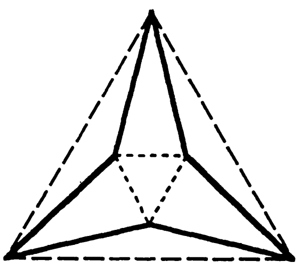
$d_1 = 8$
$d_2 = 16$
$d_3 = 9$
$e_1 = 1$
$e_2 = 0$
$e_3 = 0$



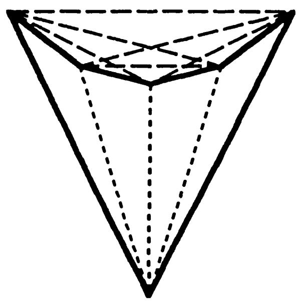
$d_1 = 4$
$d_2 = 5$
$d_3 = 2$
$e_1 = 3$
$e_2 = 2$
$e_3 = 0$



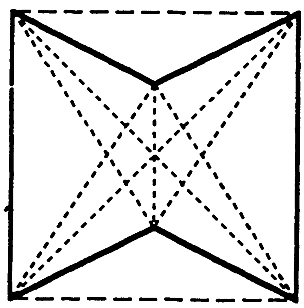
$d_1 = 6$
$d_2 = 9$
$d_3 = 4$
$e_1 = 3$
$e_2 = 2$
$e_3 = 0$



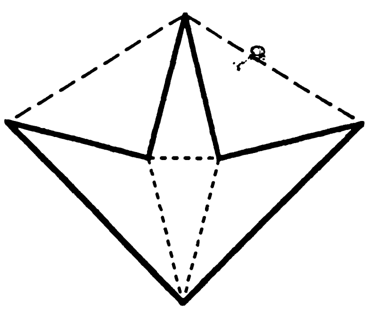
$d_1 = 3$
$d_2 = 3$
$d_3 = 1$
$e_1 = 3$
$e_2 = 3$
$e_3 = 1$



$d_1 = 3$
$d_2 = 3$
$d_3 = 1$
$e_1 = 6$
$e_2 = 10$
$e_3 = 5$



$d_1 = 7$
$d_2 = 12$
$d_3 = 6$
$e_1 = 2$
$e_2 = 1$
$e_3 = 0$



$d_1 = 3$
$d_2 = 3$
$d_3 = 1$
$e_1 = 2$
$e_2 = 1$
$e_3 = 0$

is a necessary and sufficient condition for $P(n)$ to be strictly non-convex, that is, not convex and no three consecutive vertices are collinear.

It is well-known, and easy to prove, that $n - 3 \leq d_1 \leq \frac{1}{2}n(n - 3)$, and it can be shown that every value of d_1 in this interval is the number of diagonals of some n-gon. Equality occurs on the right if $P(n)$ is strictly convex, and the inequality on

the left is a consequence of the fact that every polygon has "ears", see [2]. From these inequalities for d_1 one can deduce $0 \le e_1 \le \frac{1}{2}(n-2)(n-3)$ and every value of e_1 in this interval is the numbei of epigonals of some n-gon.

(ii) *Find the corresponding inequalities for d_i and e_i ($i = 2, 3, \ldots, n-3$). Even better, characterize all $2(n-3)$-vectors*

$$(d_1, \ldots, d_{n-3}, e_1, \ldots, e_{n-3})$$

that can arise from polygons $P(n)$ with n vertices.

ACKNOWLEDGMENT. I am indebted to Branko Grünbaum for helpful comments on an early version of this note.

REFERENCES

1. Carl W. Lee, The associahedron and triangulations of the n-gon, *Europ. J. Combinatorics* **10** (1989), pp. 551–560.
2. G. H. Meisters, Polygons have ears, *Amer. Math. Monthly* **82** (1975), pp. 648–651.

University of East Anglia
Norwich N4R 7TJ
England, UNITED KINGDOM

Every mathematician worthy of the name has experienced ... the state of lucid exaltation in which one thought succeeds another as if miraculously ... this feeling may last for hours at a time, even for days. Once you have experienced it, you are eager to repeat it but unable to do it at will, unless perhaps by dogged work.

Andre Weil, *The Apprenticeship of a Mathematician*, Birkhäuser, Boston, 1992.

It is a safe rule to apply that, when a mathematical or philosophical author writes with a misty profundity, he is talking nonsense.

Alfred North Whitehead, *An Introduction to Mathematics*, Williams and Norgate, New York, 1911.

THE AUTHORS

DANIEL HENRY GOTTLIEB received his B.S. and his Ph.D. at UCLA. The Ph.D. was granted in 1962 under the direction of S. T. Hu. He is currently a professor at Purdue University and has been an instructor at the University of Illinois and a mathematician at the Institute for Defense Analysis. His research interests center on topology and more recently in mathematical physics and mathematical history. These recent interests stem from his conviction that Topology has fundamentally changed the viewpoint of modern mathematics and that the Laws of Nature should be expressed with the concepts of topology.

ROGER HORN studied mathematics and physics as an undergraduate at Cornell University and earned his Ph.D. in mathematics at Stanford University in 1967 under the direction of Donald Spencer and Charles Loewner. He has been a faculty member at University of Santa Clara, The Johns Hopkins University (where he was chairman of the Department of Mathematical Sciences during 1972–79), University of Maryland Baltimore County, and University of Utah. His published papers and books reflect research in number theory, differential geometry, probability, health policy analysis, statistics, analysis, complex variables, and matrix analysis.

INGRAM OLKIN was an undergraduate at the College of the City of New York at a time when it produced many mathematicians. His graduate work was at Columbia University (M.S. 1949) and the University of North Carolina (Ph.D. 1951) under the direction of Harold Hotelling and S. N. Roy. He taught at Michigan State University and Minnesota before joining the Statistics Department at Stanford in 1961. His current research interests are multivariate statistical analysis, meta-analysis (combining the results of independent studies), and distribution theory.

SUSAN BASSEIN was born Richard Bassein in 1949 in New York City. As Richard, she received a PhD in Mathematics (Algebraic Geometry) in 1975 and an MS in Computer Science (Performance Evaluation) in 1982, both from UC Berkeley. She has published papers in the *Monthly* on optimal control theory, astronomy, and dynamical systems; a paper in the *Journal of Music Theory* on musical patterns; foreign language educational software; and a textbook, *An Infinite Series Approach to Calculus*. Her current passion is to explore ways to use easily and precisely controllable phenomena of electronics to illustrate the use and concepts of mathematics.

DON FALLIS received a Ph.D. in Philosophy from the University of California, Irvine, where he is now a Research Associate. His main area of research is on the epistemic status of the various methods by which mathematical knowledge is obtained (and he can't quite see what's wrong with probabilistic methods). His other interests include incompleteness, set theory, and Nietzsche's philosophy of mathematics.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Two copies of all items needed to evaluate the problem should be sent. A third copy of the the problem and solution is often useful; please include one if possible.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before November 30, 1996. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10529. *Proposed by Juan Bosco Romero Márquez, Universidad de Valladolid, Valladolid, Spain.*

Let $n \in \mathbb{Z}$, $n > 1$, and let $\lambda, a, b \in \mathbb{R}$, $\lambda \geq 0$, $0 < a \leq b$. Prove that

$$\sqrt{ab} \leq \sqrt[n]{\frac{a^n + b^n + \lambda\big((a+b)^n - a^n - b^n\big)}{2 + \lambda\big(2^n - 2\big)}} \leq \frac{a+b}{2}.$$

10530. *Proposed by Daniel Goffinet, Saint Étienne, France.*

The Cornu spiral (as a subset of the complex plane) is defined by the parameterization

$$t \mapsto z(t) = \int_0^t e^{i\pi \frac{u^2}{2}}\, du.$$

The eye sees no self-intersections. Is this a correct observation?

10531. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY, and Ieda Rodrigues, Cleveland State University, Cleveland, OH.*

Let $x > 0$. Show

$$\sum_{q=0}^{\lfloor x \rfloor} \frac{(-1)^q (x-q)^q e^{x-q}}{q!} < 2x + 1.$$

10532. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College, Bronx, NY.*

Call the n-tuple of integers $(x_1, x_2, \ldots, x_n)$ an *exceptional n-tuple* if:

$$\gcd(x_i, x_{i+1}) = 1, \text{ for } i = 1, 2, \ldots, n, \text{ with } x_{n+1} = x_1; \tag{1}$$

$$x_i \neq x_j \text{ for } i \neq j; \tag{2}$$

$$\sum_{i=1}^{n} \frac{x_i}{x_{i+1}}, \text{ with } x_{n+1} = x_1, \text{ is an integer.} \tag{3}$$

(a) For which n do there exist an infinite number of *exceptional n-tuples*?

(b) For which n does there exist an *exceptional n-tuple* in which each of the x_i is positive?

10533. *Proposed by Alfinio Flores, Arizona State University, Tempe, AZ.*

On a parallelogram P construct exterior squares on the sides. The centers of these squares form a square Q_E. On the same parallelogram construct the interior squares on the sides. The centers of these squares form another square Q_I.

(a) Show that $\text{Area}(Q_E) - \text{Area}(Q_I) = 2\,\text{Area}(P)$.

(b) Is there a generalization when P is replaced by an arbitrary convex quadrilateral?

10534. *Proposed by Paul Arne Østvær, Oslo University, Oslo, Norway.*

Suppose that R is a Noetherian ring in which all maximal ideals are principal. Show that all ideals in R are principal.

10535. *Proposed by Vladimir Janković and Jovan Vukmirović, Belgrade, Yugoslavia.*

Given s_0 with $0 < s_0 < \pi/2$, use $s_{n+1} = \sin s_n$ to define the sequence $\langle s \rangle$. Show that

$$n^2 s_n^2 - 3n + \tfrac{9}{5} \ln n$$

is convergent.

NOTES

(10534) Noetherian rings belong to the subject of *Commutative Algebra.* The first volume of the work of that title by O. Zariski & P. Samuel may be consulted for definitions and basic properties. **(10535)** In problem I.4.173 of G. Pólya and G. Szegö, *Problems and Theorems in Analysis*, the sequence $\langle s \rangle$ is shown to satisfy $\sqrt{n} s_n \to \sqrt{3}$. This refines that result.

SOLUTIONS

One Step Beyond

10288 [1993, 185]. *Proposed by Bruce R. Johnson, University of Victoria, Victoria, B. C., Canada.*

From an urn containing b balls, numbered from 1 to b, balls are drawn one at a time with replacement until the accumulated sum of all numbers drawn is at least equal to a positive integer n. Let X_n denote the amount by which the accumulated sum exceeds n. Find $\lim \mathbf{E}(X_n)$ or show that this limit does not exist.

Solution I by David Callan, University of Wisconsin, Madison, WI. For $0 \leq k < b$, the event $X_n = k$ is a disjoint union of the compound events: (1) the accumulated sum hits $n - i$ (denote the probability of this by p_{n-i}); (2) the next draw is $i + k$. These conditions require $1 \leq i \leq b - k$. Hence

$$\mathbf{E}(X_n) = \sum_{k=0}^{b-1} k \sum_{i=1}^{b-k} p_{n-i}\frac{1}{b} = \frac{1}{b}\sum_{i=1}^{b} p_{n-i}\frac{(b-i)(b-i+1)}{2}.$$

It was shown in *Math. Magazine* Problem 1217 [1985, 177; 1986, 174] that, for large n, $p_n \to 2 \,/\, (b+1)$. Hence $\lim \mathbf{E}(X_n)$ exists, and is easily computed to be $(b-1)/3$.

Solution II by Uwe Jensen, University of Ulm, Ulm, Germany. More generally, suppose the balls are numbered with nonnegative integers whose greatest common divisor is 1. Let Y_i be the number of the ball in the i^{th} drawing, $\mu = \mathbf{E}\,Y_i$, $\mu_2 = \mathbf{E}\,Y_i^2$, $i = 1, 2, \ldots$; $S_k = \sum_{i=1}^{k} Y_i$; $N_n = \max\{k \in \mathbb{N} : S_k \leq n\}$ (where we take $\max\emptyset = 0$). Then N_n is known as a *renewal process*, and $X_n = S_{N_{n-1}+1} - n$ is called the *excess life* or *forward recurrence time*. Wald's identity gives

$$\mathbf{E}\,X_n = \mu\,\mathbf{E}(N_{n-1} + 1) - n.$$

Using the asymptotic expression from W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I (third edition), Wiley, 1968, p. 341,

$$\mathbf{E}\,N_n = \frac{n}{\mu} + \frac{\mu_2 + \mu}{2\mu^2} - 1 + \epsilon_n$$

with $\epsilon_n \to 0$ as $n \to \infty$, we get

$$\lim_{n\to\infty} \mathbf{E}\,X_n = \frac{\mu_2}{2\mu} - \frac{1}{2}.$$

(The condition on the greatest common divisor is used in the results quoted here.) In the case in which Y_i is uniformly distributed on $\{1, 2, \ldots, b\}$, it follows that

$$\mu = \sum_{i=1}^{b}\left(\frac{1}{b}\right)i = \frac{b+1}{2}$$

$$\mu_2 = \sum_{i=1}^{b}\left(\frac{1}{b}\right)i^2 = \frac{(2b+1)(b+1)}{6}$$

$$\lim_{n\to\infty} \mathbf{E}\,X_n = \frac{2b+1}{6} - \frac{1}{2} = \frac{b-1}{3}.$$

In fact, in this uniform distribution case, the formula

$$\mathbf{E}(N_{n-1}+1)=\sum_{v=0}^{\infty}(-1)^v\binom{n-1-bv}{v}\frac{1}{b^v}\left(1+\frac{1}{b}\right)^{n-1-(b+1)v}$$

was obtained in U. Jensen, "Some remarks on the renewal function of the uniform distribution", *Adv. Appl. Prob.* 16 (1984), 214–215, allowing the individual $\mathbf{E}\,X_n$ to be found as well as the limiting value.

Solved also by R. A. Agnew, M. H. Andreoli, R. J. Chapman (U. K.), C. Cooper & R. E. Kennedy & A. Tinsley & H. Chen & L. Cammack, D. A. Darling, R. L. Doucette, H. von Eitzen (Germany), P. Griffin, E. Hertz, R. Holzsager, I. Kastanas, H. G. Killingbergtrø (Norway), J. H. Lindsey II, O. P. Lossers (The Netherlands), A. Pedersen (Denmark), B. Peterson, R. M. Robinson, I. Skau (Norway), M. Stamp, J. H. Steelman, H. L. Stubbs, A. A. Tarabay (Lebanon), D. Wolfe, A. N. 't Woord (The Netherlands), H. Zeisel (Austria), P. J. Zwier, GCHQ Problem Solving Group (U. K.) (two solutions), Western Maryland College Problems group, and the proposer. One incorrect and three incomplete solutions were received.

A Recurring Integral

10309 [1993, 499]. *Proposed by Walter Rudin, University of Wisconsin, Madison, WI.*

Compute

$$\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log(A+B\cos\theta)\,d\theta\right)$$

when $A>B>0$. The answer should be given as an algebraic function of A and B.

Solution I by Nicholas Passell, University of Wisconsin, Eau Claire, WI. Let $I=\int_{-\pi}^{\pi}\log(A+B\cos\theta)\,d\theta$. The integrand is periodic with period 2π, so the substitution $\theta=\phi+\pi$ shows that the value of I does not depend on the sign of B. Adding together the integrals for B and $-B$, one gets $2I=\int_{-\pi}^{\pi}\log(A^2-B^2\cos^2\theta)\,d\theta$. Replace $\cos^2\theta$ by $(\cos 2\theta+1)\ /\ 2$, make the change of variables $\phi=2\theta$, and again use the periodicity of the integrand to obtain

$$2I=\int_{-\pi}^{\pi}\log\left(A^2-\frac{B^2}{2}-\frac{B^2}{2}\cos\phi\right)d\phi.$$

To exploit this transformation, write $|B|=2C$ and $\alpha=\left(A+\sqrt{A^2-B^2}\right)\ /\ |B|$. Then, the two formulas can be written:

$$I=2\pi\log C+\int_{-\pi}^{\pi}\log\left(\alpha+\alpha^{-1}-2\cos\theta\right)d\theta;$$

$$2I=2\pi\log\left(C^2\right)+\int_{-\pi}^{\pi}\log\left(\alpha^2+\alpha^{-2}-2\cos\theta\right)d\theta.$$

Iterating the construction thus gives

$$I=2\pi\log C+\frac{1}{2^k}\int_{-\pi}^{\pi}\log\left(\alpha^{2^k}+\alpha^{-2^k}-2\cos\theta\right)d\theta.$$

As $k\to\infty$, this gives $I=2\pi\log(C\alpha)$. The original quantity is thus

$$C\alpha=\frac{A+\sqrt{A^2-B^2}}{2}.$$

Solution II by Sayel Ali, Moorhead State University, Moorhead, MN. Suppose that $A>B>0$. Let

$$r=\frac{\sqrt{A+B}-\sqrt{A-B}}{2}\quad\text{and}\quad a=\frac{\sqrt{A+B}+\sqrt{A-B}}{2}.$$

Let $\overline{B}(a; r)$ be the closed disc in the complex plane with center on the x-axis at a and radius r. Then $\overline{B}(a; r) \subseteq \mathbb{C} - \{0\}$. Since $u(z) = \log|z|^2$ is harmonic on $\mathbb{C} - \{0\}$, applying the mean value theorem (see J. B. Conway, *Functions of One Complex Variable*, Second Edition, Springer-Verlag, 1978, p. 253) to $u(z)$ on the circle $|z - a| = r$ leads to

$$\log\left(a^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|a + re^{i\theta}\right|^2 d\theta.$$

Now,

$$\left|a + re^{i\theta}\right|^2 = a^2 + r^2 + 2ar\cos\theta = A + B\cos\theta,$$

so that

$$\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(A + B\cos\theta\right) d\theta\right) = a^2 = \frac{A + \sqrt{A^2 - B^2}}{2}.$$

Editorial comment. Other methods of finding the integral involved writing

$$\log\left(1 + r\cos\theta\right) = \sum_{n=1}^{\infty}(-1)^n \frac{r^n \cos^n\theta}{n},$$

and integrating the series term-by-term. The resulting series is recognized by exploiting its relation with the series representation of $(1 - x^2)^{-1/2}$.

Another approach identified $f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(x + \cos\theta)\, d\theta$ by differentiating with respect to x. Methods of elementary calculus allow $f'(x)$ to be found and integrated. The determination of $f(x)$ is completed by evaluating $f(1)$. This can be done by using half angle formulas to relate $f(1)$ to the known integral $\int_0^{\pi/2} \log\sin\theta\, d\theta = -(\pi/2)\log 2$. Alternatively, as noted by David Cruz-Uribe, this determines the integral I of solution I as a function of A and B up to an additive constant. The constant is easily found by setting $B = 0$ in this formula.

Symbolic packages do not yet know the value of this integral, but S. C. Locke was able to use *Maple* to find several terms of the series in B for the desired expression. The answer could then be guessed. Any of the above methods could then verify that the guess was correct, with verification being easier than discovery.

Several readers were more skilled than the editors in finding this integral in tables of integrals. Entry 2.6.36.29 on p. 546 of A. P. Prudnikov et al., *Integrals and Series*, Vol. 1, Gordon and Breach, 1986 gives the integral in the more general form with $B\cos\theta$ replaced by $B_0\cos\theta + B_1\sin\theta$ with $B_0^2 + B_1^2 = B^2$. This additional generality may be obtained from the special case by a suitable translation of θ as at the beginning of Solution I.

Nelson M. Blachman reported encountering a similar integral in an engineering application. See N. M. Blachman, "Phase and logarithm of amplitude of a signal plus noise", *Electron. Lett.*, 24 (1988), 482–483 for details.

Solved by 81 readers (including those cited) and the proposer. Two of these submitted more than one solution. There were also two incorrect solutions.

Three of a Kind

10320 [1993, 590]. *Proposed by Ignacy I. Kotlarski, Oklahoma State University, Stillwater, OK.*

Under the assumption that f_0, f_1 and f_2 are defined on $[0, \infty)$, Laplace transformable, and not equivalent to zero, solve the integral equation

$$\int_0^{\min(x_1,x_2)} f_0(x)\, f_1(x_1 - x)\, f_2(x_2 - x)\, dx = e^{-\max(x_1,x_2)}\left(1 - e^{-\min(x_1,x_2)}\right),$$

with $x_1 \geq 0$ and $x_2 \geq 0$, for the three functions f_0, f_1 and f_2.

Solution by Kenneth F. Andersen, University of Alberta, Edmonton, Alberta, Canada. An easy calculation shows that the equation is satisfied if $f_j(x) = c_j e^{-x}$ for constants c_j, $j = 0, 1, 2$ satisfying $c_0 c_1 c_2 = 1$. We shall show that these are the only solutions.

Let F_j denote the Laplace transform of f_j. By hypothesis, each $F_j(s)$ is defined and continuous for all large s, say $s \geq \sigma \geq 0$, and

$$\lim_{s\to\infty} F_j(s) = 0, \quad j = 0, 1, 2. \tag{1}$$

The change of variables $t_j = x_j - x$ shows

$$\begin{aligned} &F_0(s_1+s_2)\,F_1(s_1)\,F_2(s_2) = \\ &\quad = \int_0^\infty e^{-(s_1+s_2)x} f_0(x)\left(\int_0^\infty e^{-s_1t_1} f_1(t_1)\,dt_1 \int_0^\infty e^{-s_2t_2} f_2(t_2)\,dt_2\right)dx \\ &\quad = \int_0^\infty f_0(x)\left(\int_x^\infty e^{-s_1x_1} f_1(x_1-x)\,dx_1 \int_x^\infty e^{-s_2x_2} f_2(x_2-x)\,dx_2\right)dx; \end{aligned}$$

and Fubini's theorem shows that this equals

$$\int_0^\infty \int_0^\infty e^{-s_1x_1-s_2x_2}\left(\int_0^{\min(x_1,x_2)} f_0(x)\,f_1(x_1-x)\,f_2(x_2-x)\,dx\right)dx_1\,dx_2.$$

On the other hand, an easy calculation shows

$$\begin{aligned} &\int_0^\infty \int_0^\infty e^{-s_1x_1-s_2x_2} e^{-\max(x_1,x_2)}\left(1-e^{-\min(x_1,x_2)}\right)dx_1\,dx_2 = \\ &\qquad = \frac{1}{(1+s_1+s_2)(1+s_1)(1+s_2)}. \end{aligned}$$

Hence, the given equation shows that we must have

$$F_0(s_1+s_2)\,F_1(s_1)\,F_2(s_2) = \frac{1}{(1+s_1+s_2)(1+s_1)(1+s_2)}, \quad s_1, s_2 \geq \sigma. \tag{2}$$

For $s > 0$, set

$$G_0(s) = (1+2\sigma+s)F_0(2\sigma+s)$$

and

$$G_j(s) = \left((1+\sigma+s)F_j(\sigma+s)\right)^{-1}$$

for $j = 1, 2$. Then, for $s_1, s_2 \geq 0$, (2) shows $G_1(s_1)G_2(s_2) = G_0(s_1+s_2) = G_0(s_2+s_1) = G_1(s_2)G_2(s_1)$, which implies that $G_2(s) = cG_1(s)$ for $s \geq 0$ with $c = G_2(0) \,/\, G_1(0)$. Thus, (2) becomes

$$G_0(s_1+s_2) = cG_1(s_1)G_1(s_2) \tag{3}$$

or, setting $H(s) = G_0(s) \,/\, G_0(0)$, H satisfies

$$H(s_1+s_2) = H(s_1)H(s_2).$$

It is well known that this implies $H(s) = A^s$ for some $A > 0$. Thus, $G_0(s) = G_0(0)\,A^s$ and (3) yields $G_1(s) = \pm\sqrt{G_0(0)/c}\,A^s$. Then

$$F_0(2\sigma+s) = \frac{G_0(0)A^s}{1+2\sigma+s}$$

and (1) shows that we must have $A \leq 1$. On the other hand,

$$F_1(\sigma+s) = \frac{\pm\sqrt{c \,/\, G_0(0)}\,A^{-s}}{1+\sigma+s}$$

and(1) shows that $A \geq 1$. Thus $A = 1$ and, therefore, $F_j(s) = c_j \,/\, (1+s)$ with $c_0 = G_0(0)$, $c_1 = \pm\sqrt{c/G_0(0)}$, $c_2 = \pm 1 \,/\, \sqrt{cG_0(0)}$. Hence, the $f_j(x)$ are as claimed.

Solved also by R. D. Brown & P. Szeptycki, R. J. Chapman (U. K.), D. A. Darling, D. K. Nester, D. Schepler (student), A. N. 't Woord (The Netherlands), and the proposer.

Scaling Sums into Integrals

10321 [1993, 590]. *Proposed by Carl Axness, Sandia National Laboratories, Albuquerque, NM, Reinhard Schäfke, University of Essen, Essen, Germany, and David Arterburn, New Mexico Tech, Socorro, NM.*

Let μ be a positive real number. Prove

$$\lim_{x\to 1^+} (\ln x)^{1/\mu} \sum_{i=1}^{\infty} x^{-(2i-1)^{\mu}} = \frac{\Gamma(1/\mu)}{2\mu}.$$

Solution by Heinz-Jürgen Seiffert, Berlin, Germany. First we prove the following

Theorem. *If the function $f : \mathbb{R}^{\geq 0} \to \mathbb{R}$ is nonnegative and nonincreasing, then for all real numbers a and b with $a > \max(b, 0)$,*

$$\lim_{h\to 0^+} h \sum_{i=1}^{\infty} f\big((ai-b)h\big) = \frac{1}{a}\int_0^{\infty} f(t)\,dt, \tag{1}$$

provided that the integral exists.

Proof. There is no loss of generality in letting $a = 1$, since the general case may be recovered by replacing h by ha and b by b/a. For $h > 0$, let $x_i = (i-b)h$, $i \in \mathbb{N}$. Since f is nonincreasing, it follows that

$$hf(x_{i+1}) \leq \int_{x_i}^{x_{i+1}} f(t)\,dt \leq hf(x_i),$$

for all $i \in \mathbb{N}$. Summing gives

$$h\sum_{i=2}^{n} f(x_i) \leq \int_{x_1}^{x_n} f(t)\,dt \leq h\sum_{i=1}^{n-1} f(x_i).$$

Since f is nonnegative and the integral exists, we may let n tend to infinity to obtain

$$h\sum_{i=1}^{\infty} f(x_i) - hf(x_1) \leq \int_{x_1}^{\infty} f(t)\,dt \leq h\sum_{i=1}^{\infty} f(x_i).$$

Now, the result obviously follows.

This generalizes the case where $a = 1$ and $b = 0$, which is in G. Pólya and G. Szegö, *Problems and Theorems in Analysis*, Vol. I, Springer-Verlag, 1970, p. 53, entry 30.

The conditions of the above theorem are satisfied for the function $f(t) = e^{-t^{\mu}}$, $t \geq 0$. Hence, we get

$$\lim_{h\to 0^+} h\sum_{i=1}^{\infty} e^{-(ai-b)^{\mu}h^{\mu}} = \frac{1}{a}\int_0^{\infty} e^{-t^{\mu}}\,dt.$$

Substituting $h = (\ln x)^{1/\mu}$, $x > 1$, and using

$$\int_0^{\infty} e^{-t^{\mu}}\,dt = \frac{1}{\mu}\int_0^{\infty} y^{\frac{1}{\mu}-1}e^{-y}\,dy = \frac{\Gamma(1/\mu)}{\mu},$$

we find

$$\lim_{x\to 1^+} (\ln x)^{1/\mu} \sum_{i=1}^{\infty} x^{-(ai-b)^{\mu}} = \frac{\Gamma(1/\mu)}{a\mu}.$$

The stated problem is the special case where $a = 2$ and $b = 1$.

Editorial comment. In a second solution, H.-J. Seiffert quoted

Theorem. *Let $\langle a\rangle$ be a sequence of nonnegative real numbers such that $\lim_{n\to\infty} a_n/n = k$, where $k > 0$. Also let $f : \mathbb{R}^{\geq 0} \to \mathbb{R}$ be nonnegative and nonincreasing such that $\int_0^{\infty} f(t)\,dt$ exists. Then*

$$\lim_{h\to 0^+} h \sum_{i=1}^{\infty} f(ha_i) = \frac{1}{k}\int_0^{\infty} f(t)\,dt.$$

Proof. See J. J. A. M. Brands, "Problem 882", Nieuw Archief voor Wiskunde (Ser. 4) 2 (1993), p. 175.

The theorem in the selected solution is clearly a special case of this result.

Martin Engman gave a solution similar to the selected solution, and claimed P. Gilkey, *Invariance Theory, the Heat Equation, and the Atiyah-Singer Index Theorem*, Publish or Perish, 1984, p. 317 as his inspiration, where a form of (1) with $f(x) = e^{-x^2}$ is studied.

S. K. Rangarajan, T. P. T. Williams, the MMRS group, and the proposers (in a second solution) used the theory of Mellin transforms to give additional terms in an asymptotic series for the particular function of x appearing here.

Solved also by K. F. Andersen (Canada), J. Anglesio (France), G. Bach (Germany), W. Blumberg, R. J. Chapman (U. K.), D. A. Darling, M. Engman, I. Kastanas, S. Koumandos (Australia), K.-W. Lau (Hong Kong), J. H. Lee (student, Korea), O. P. Lossers (The Netherlands), D. K. Nester, A. Nijenhuis, S. K. Rangarajan (India), A. Sinefakopoulos (student, Greece), T. P. T. Williams (U. K.), the MMRS group of Oklahoma State University, and the proposers. One incomplete solution was received.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, T?(14–15), S. *Mathematik für Wirtschaftswissenschaftler.* Volker Nollau. BG Teubner Leipzig, 1995, 264 pp, DM 29,80 (P). [ISBN 3-8154-2049-0] Second, revised and slightly expanded edition. (*First Edition*, TR, August–September 1994.) JD-B

General, S(13). *Starthilfe Mathematik.* Winfried Schirotzek, Siegfried Scholz. Mathematik für Ingenieure und Naturwissenschaftler. BG Teubner Leipzig, 1995, 139 pp, DM 19,80 (P). [ISBN 3-8154-2085-7] On the school mathematics assumed in German university science and engineering courses. Chapters on logic and set theory, number systems, functions of a real variable, elementary functions, vectors, plane geometry, linear equations, sequences, limits and continuity, differential calculus, integral calculus. JD-B

General, P. *The Floer Memorial Volume.* Eds: Helmut Hofer, *et al.* Progress in Math., V. 133. Birkhäuser Boston, 1995, xii + 685 pp, $98. [ISBN 0-8176-5044-8] 28 papers on gauge theory, dynamical systems, symplectic geometry, and topology.

Reference, S(13). *All You Wanted to Know About Mathematics But Were Afraid to Ask: Mathematics for Science Students, Volume 1.* Louis Lyons. Cambridge Univ Pr, 1995, xviii + 325 pp, $24.95 (P); $59.95. [ISBN 0-521-43600-1; 0-521-43465-3] Despite the title, not a "math for poets" book. It is, rather, a good reference volume for undergraduates that serves as a guide to simultaneous equations, geometry, vectors, complex numbers (a particularly nice treatment including applications in electrical circuits with resonance), ordinary differential equations (including a treatment of the D operator), partial derivatives, Taylor series, and Langrange multipliers. Numerous interesting applications. KS

Mathematics Appreciation, S(13–16), L*. *200% of Nothing.* A.K. Dewdney. Wiley, 1993, ix + 182 pp, $12.95 (P). [ISBN 0-471-14574-2] Paperback reprint of a witty and feisty collection of distortions, exaggerations, misrepresentations, and perversions that advertisers, journalists, politicians, and merchants foist on an innumerate public. Great source of motivational examples for introductory mathematics and statistics courses. LAS

Finite Mathematics, T(13–14). *Finite Mathematics: An Applied Approach, Seventh Edition.* Abe Mizrahi, Michael Sullivan. Wiley, 1996, xix + 691 pp. [ISBN 0-471-10700-X] Features exercises requiring graphing calculator or software. Questions from CPA, CMA, and actuarial exams included where feasible. New edition extensively rewritten for clarity and to update material; reorganized to allow multiple paths through material. (*Fifth Edition*, TR, January 1989.) LB

Education, P, L. *Abstraction and Representation: Essays on the Cultural Evolution of Thinking.* Peter Damerow. Stud. in the Phil. of Sci., V. 175. Kluwer Academic, 1996, xiv + 414 pp, $170. [ISBN 0-7923-3816-2] English translations of selected papers (originally published in German) dealing with the evolution of mathematical thinking from pre-literate societies to contemporary mathematics education. Offers

probing insights on the nature of abstract thinking and the role of mathematics education based on broad cultural and historical research. LAS

Education, L*. *Learn From the Masters!* Eds: Frank Swetz, *et al.* MAA, 1995, x + 303 pp, $23 (P). [ISBN 0-88385-703-0] Articles on how to incorporate history into the teaching of high school or college mathematics. Articles cover general ideas on how history can be used, as well as specific topics (for example, logarithms, vectors). Should be of interest to any teacher who wishes to enrich his/her class with historical material. LC

History, L.** *Ramanujan: Letters and Commentary.* Bruce C. Berndt, Robert A. Rankin. History of Math., V. 9. AMS, 1995, xiv + 347 pp, $79; $49 (P). [ISBN 0-8218-0470-7; 0-8218-0287-9] Collection of letters written by, or to, or about S. Ramanujan, including the first letters Ramanujan wrote to G.H. Hardy, and the latter's response. Commentaries accompany each letter providing biographical sketches, mathematical notes, cultural background, etc. LC

Logic, P. *Logic and Visual Information.* Eric M. Hammer. Stud. in Logic, Lang. & Inform. Ser. Center for Study of Language & Information (Stanford U., Ventura Hall, Stanford, CA 94305), 1995, ix + 124 pp, $18.95 (P); $49.95. [ISBN 1-881526-99-2; 1-881526-87-9] Logical foundations of information presented in visual forms, e.g., diagrams, graphs, maps, charts. Discusses relationship between language and visually-presented information. Develops formalization of reasoning using several types of diagrams (diagrammatic logics), and shows these systems to be logically sound and complete. LB

Foundations, P. *Set Theory.* Eds: Tomek Bartoszyński, Marion Scheepers. Contemp. Math., V. 192. AMS, 1996, xii + 184 pp, $45 (P). [ISBN 0-8218-0306-9] Papers from the first three meetings (in 1992, 1993, and 1994) of the annual Boise Extravaganza in Set Theory held at Boise State University.

Number Theory, T(16–17: 1, 2). *A Course in Number Theory, Second Edition.* H.E. Rose. Oxford Univ Pr, 1995, xv + 398 pp, $35 (P). [ISBN 0-19-852376-9] This edition has a second chapter on elliptic curves, new material on factorization, primality testing, and the RSA public-key cryptosystem. An extremely demanding text for undergraduates, but well-suited for a mathematician who wants to learn some number theory. Comparable to Hardy and Wright's classic text, but far more current. (1988 edition, TR, February 1989.) DB

Linear Algebra, T(14–16: 1, 2), L. *Linear Algebra.* Terry Lawson. Wiley, 1996, xvi + 408 pp, $67.95. [ISBN 0-471-30897-8] More sophisticated treatment than most recent elementary texts. Pace is brisk. Covers determinants, LU-decomposition, vector spaces, linear transformations in the first two chapters. Succeeding chapters on orthogonality and projections, diagonalization, the spectral theorem, normal forms. Numerous applications, computer computations, exercises of graded difficulty. JS

Linear Algebra, S(14), P. *Matrix Algebra As A Tool.* Ali S. Hadi. Duxbury Pr, 1996, xi + 212 pp, $25.75 (P). [ISBN 0-534-23712-6] Informal presentation of terminology and mechanics of matrix algebra. Covers matrix arithmetic, reduction, vector geometry, and introduces several applications, including descriptive statistics and regression analysis. No computer usage. JNC

Linear Algebra, P. *Matrices of Sign-solvable Linear Systems.* Richard A. Brualdi, Bryan L. Shader. Tracts in Math., V. 116. Cambridge Univ Pr, 1995, xii + 298 pp, $49.95. [ISBN 0-521-48296-8] A linear system $A\mathbf{x} = \mathbf{b}$ is sign-solvable if the signs of the entries of $\mathbf{x}$ can be determined from the signs of the entries in A and $\mathbf{b}$. Studies properties of matrices that can be deduced from the arrangement of the positive, negative, and zero entries. Intended primarily for researchers in combinatorics and linear algebra. LC

Group Theory, P. *Moonshine, the Monster, and Related Topics.* Eds: Chongying Dong, Geoffrey Mason. Contemp. Math., V. 193. AMS, 1996, ix + 368 pp, $70 (P). [ISBN 0-8218-0385-9] Proceedings of a 1994 AMS–IMS–SIAM Joint Summer Research Conference held at Mt. Holyoke College.

Group Theory, P. *Groups—Korea '94.* Eds: A.C. Kim, D.L. Johnson. Walter de Gruyter, 1995, ix + 344 pp, DM 198. [ISBN 3-11-014793-9] Proceedings of the Third International Conference on the Theory of Groups held in Pusan in 1994. Main emphasis is on geometric group theory.

Algebra, T(15–16: 1), L. *Elements of Modern Algebra, Fourth Edition.* Jimmie Gilbert, Linda Gilbert. PWS, 1996, xi + 372 pp, $67.50. [ISBN 0-534-95196-1] Self-contained introduction to basic algebra (groups, rings, polynomials), oriented towards prospective secondary math teachers. Lots of routine exercises, optional sections on coding theory, geometry and art, biographical sketches of contributors to the field. (*Third Edition*, TR, June–July 1992). RM

Algebra, T(17–18: 1), S, P. *Quadratic Forms with Applications to Algebraic Geometry and Topology.* Albrecht Pfister. London Math. Soc. Lect. Note Ser., V. 217. Cambridge Univ Pr, 1995, viii + 179 pp, $34.95 (P). [ISBN 0-521-46755-1] Nice introduction to quadratic forms, flavored by the author's 30 years of contributions. Themes include field invariants, Hilbert's 17th problem, interconnections with algebraic geometry and topology. RM

Algebra, T(16–17: 1), S, P, L. *Clifford Algebras and the Classical Groups.* Ian R. Porteous. Stud. in Adv. Math., V. 50. Cambridge Univ Pr, 1995, x + 295 pp, $49.95. [ISBN 0-521-55177-3] In response to increased interest in physics in Clifford algebras, the author has re-worked his topological geometry and expanded the Clifford algebra material to include a full and largely self-contained treatment and relation to linear groups. Assumes some linear algebra and calculus, but suitable for advanced undergraduates. JS

Calculus, T(13). *Calculus: Modeling and Application.* David A. Smith, Lawrence C. Moore. DC Heath, 1996, xviii + 714 pp, (P). [ISBN 0-669-32787-5] Product of the Duke calculus reform project. Emphasizes "discovery learning" by developing topics through a series of explorations, checkpoints, and exercises. Differential equations in real-world contexts motivate single variable calculus concepts. Contains many exercises and projects intended for small groups. Requires a graphing calculator or software. Laboratory manuals available for some graphing calculators and computer algebra systems. TR

Calculus, S(13–14). *Analysis in Fragen und Übungsaufgaben.* Karl-Heinz Gärtner, *et al.* Mathematik für Ingenieure und Naturwissenschaftler. BG Teubner Leipzig, 1995, 264 pp, DM 26,80 (P). [ISBN 3-8154-2088-1] Definitions, theorems, questions and answers, exercises and solutions. Intended for use with course or to prepare for examination. JD-B

Partial Differential Equations, T(18: 2), L. *Introduction to Partial Differential Equations, Second Edition.* Gerald B. Folland. Princeton Univ Pr, 1995, xi + 324 pp, $39.50. [ISBN 0-691-04361-2] Originally published in the Princeton series Mathematical Notes (TR, January 1977). Revisions include additional material (including a new chapter on pseudodifferential operators), improved exposition, and more exercises. LC

Dynamical Systems, P. *Dynamical Systems and Probabilistic Methods in Partial Differential Equations.* Eds: Percy Deift, C. David Levermore, C. Eugene Wayne. Lect. in Appl. Math., V. 31. AMS, 1996, ix + 268 pp, $29 (P). [ISBN 0-8218-0368-9] Proceedings of the 1994 AMS–SIAM Summer Seminar on Dynamical Systems and Probabilistic Methods for Nonlinear Waves held at MSRI.

Numerical Analysis, P, L. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* K.E. Brenan, S.L. Campbell, L.R. Petzold. SIAM, 1996, x + 256 pp, $29.50 (P). [ISBN 0-89871-353-6] Unabridged, corrected republication of the 1989 North-Holland edition. A new chapter summarizes significant developments since the book was originally published, and updates the bibliography.

Analysis, T(17: 1), L. *Polynomials and Polynomial Inequalities.* Peter Borwein, Tamás Erdélyi. Grad. Texts in Math., V. 161. Springer-Verlag, 1995, x + 480 pp, $59. [ISBN 0-387-94509-1] Analytic properties of polynomials and rational functions of one variable. Studies special polynomials (e.g., Chebyshev), polynomial inequalities, questions of density, etc. Substantial number of exercises. Tersely written. Assumes undergraduate-level real and complex analysis. Appropriate for self-study. LC

Analysis, T(17: 2). *Modern Analysis and Topology.* Norman R. Howes. Universitext. Springer-Verlag, 1995, xxviii + 403 pp, $39 (P). [ISBN 0-387-97986-7] Integrated approach to analysis and topology via uniform spaces. (Uniform spaces are more general than metric spaces, but provide more structure than arbitrary topological spaces.) Begins with first principles of uniform spaces, and concludes with results from modern analysis proved in the last decade. TR

Differential Geometry, T?(17–18: 1), S, P. *An Introduction to Noncommutative Differential Geometry and its Physical Applications.* J. Madore. London Math. Soc. Lect. Note Ser., V. 206. Cambridge Univ Pr, 1995, 200 pp, $32.95 (P). [ISBN 0-521-46791-8] For a "space" V, a geometric structure is usually reflected in some nice (commutative!) algebra $C(V)$ of complex valued functions on V, and recoverable from $C(V)$. The goal of noncommutative geometry is to reformulate the notion of geometry for non-commutative algebras, for which there is no natural set of "points." Crisp and readable introduction, concentrating on the matrix and finite dimensional case for noncommutative differential geometry, nicely motivated by applications to physics. RM

Differential Geometry, T(15: 1), S. *Tensorrechnung.* Hans Karl Iben. Mathematik

für Ingenieure und Naturwissenschaftler. BG Teubner Leipzig, 1995, 180 pp, DM 22,80 (P). [ISBN 3-8154-2083-0] Introduction to tensors, intended chiefly for engineers, scientists. Some exercises, hints on solutions. JD-B

Geometry, T(16–17), L. *Combinatorial Geometry.* János Pach, Pankaj K. Agarwal. Ser. in Disc. Math. & Optim. Wiley, 1995, xiii + 354 pp, $59.95. [ISBN 0-471-58890-3] Studies problems about arrangements of points, lines, and convex bodies—for example, how many unit balls can be placed in a box of fixed volume. Material has ties to problems in coding theory, computer graphics, computational geometry. Text assumes calculus and some basic probability and combinatorics. Includes exercises, extensive bibliography. LC

Geometry, T(15: 1), S, P. *Projective Geometry and Modern Algebra.* Lars Kadison, Matthias T. Kromann. Birkhäuser Boston, 1996, xvi + 208 pp, $44.50. [ISBN 0-8176-3900-4] Uses synthetic and analytic approaches to explore properties of the real projective plane and more general projective planes over other fields and division rings. Group theory is introduced to facilitate the focus on automorphism groups of these planes. JNC

Geometry, T*(15: 1), P, L. *Fundamentals of Modern Elementary Geometry.* Howard Eves. Jones & Bartlett, 1992, x + 198 pp. [ISBN 0-86720-247-5] Material from Volume One of the author's *A Survey of Geometry* (Allyn and Bacon, 1963, revised 1972) repackaged for preservice and in-service high school teachers. The three chapters (Modern Elementary Geometry, Elementary Transformations, Euclidean Constructions) make some of Eves' classic work available again. JNC

Geometry, T(13: 1). *Introduction to Geometry.* Marjorie Anne Fitting. McGraw-Hill, 1996, xiii + 476 pp, $38 (P). [ISBN 0-07-021182-5] A problem-solving, activity approach to elementary Euclidean geometry. Appendices include work with Turtle Graphics, *The Geometer's Sketchpad*, and Miras. Appropriate for pre-calculus students or pre-service elementary teachers with previous geometry experience. JNC

Algebraic Topology, T(16: 2). *Algebraic Topology: A First Course.* William Fulton. Grad. Texts in Math., V. 153. Springer-Verlag, 1995, xviii + 430 pp, $59.50. [ISBN 0-387-94326-9] Emphasizes concrete problems in low dimensions, and the connection between topology and analysis (e.g., the relationship between homology and integration). Homology, cohomology, and homotopy are first introduced for planar regions, and then extended to surfaces with some indication for extending to higher dimensions. Many exercises and challenging problems. TR

Topology, P. *The Interface of Knots and Physics.* Ed: Louis H. Kauffman. Proc. of Symp. in Appl. Math., V. 51. AMS, 1996, x + 208 pp, $39. [ISBN 0-8218-0380-8] Lecture notes prepared for the January 1995 AMS Short Course held in San Francisco.

Optimization, T(14–17: 1, 2). *Variational Calculus and Optimal Control: Optimization with Elementary Convexity, Second Edition.* John L. Troutman. Undergrad. Texts in Math. Springer-Verlag, 1996, xv + 461 pp, $49.95. [ISBN 0-387-94511-3] Broad-ranging introduction to optimization, from classic time-of-transit problems to modern optimal control theory. Many examples and exercises. (*First Edition*, January 1984.) DB

Mathematical Modeling, S?(17–18), P?, L. *Mathematical Exploration of the Environment.* István Bán. Transl: Ákos Wallner. Akadémiai Kiadó, 1995, 116 pp, $20. [ISBN 963-05-6820-9] Introduces new concept of environment structure—porous structure—which applies sets of state characteristic values. Provides mathematical method to process a large number and different kinds of characteristic values, surveys the complex correlations, and explains local and global properties of the environment by means of different mathematical interpretations. Describes planned method of selection by Bán (PMSB) in mathematical detail which can be applied to economic policy, public health, industry, etc. KB

Optimal Control, P. *Lecture Notes in Control and Information Sciences–207: Optimal Feedback Control.* R. Gabasov, F.M. Kirillova, S.V. Prischepova. Springer-Verlag, 1995, xv + 202 pp, $54 (P). [ISBN 3-540-19991-8]

Optimal Control, P. *Exact Controllability and Stabilization: The Multiplier Method.* V. Komornik. Res. in Appl. Math. Wiley, 1994, viii + 156 pp, $39.95 (P). [ISBN 0-471-95367-9]

Probability, S, L*. *A Philosophical Essay on Probabilities.* Pierre Simon, Marquis de Laplace. Dover, 1995, viii + 196 pp, $7.95 (P). [ISBN 0-486-28875-7] Unaltered and unabridged republication of the original 1902 Wiley publication, previously reprinted by Dover in 1952. Written as a popular introduction to his *Théorie analytique des probabilités*, this classic presents concepts and applications of probability in nonmathematical terms. Source of the often quoted and variously

translated "... the most important questions of life, which are indeed for the most part only problems of probability." RSK

Stochastic Processes, P. *Dirichlet Forms and Stochastic Processes.* Eds: Z.M. Ma, M. Röckner, J.A. Yan. Walter de Gruyter, 1995, xi + 443 pp, DM 268. [ISBN 3-11-014284-8] 38 papers from a 1993 International Conference held in Beijing.

Elementary Statistics, T(13), C. *A Casebook for a First Course in Statistics and Data Analysis.* Samprit Chatterjee, Mark S. Handcock, Jeffrey S. Simonoff. Wiley, 1995, xi + 314 pp, $24.95 (P), with disk. [ISBN 0-471-11030-2] Nice, diverse set of introductory statistical applications. All analyses motivated by scientific questions. Some analyses carried out completely, some indicate how one might proceed, and some are left for students to analyze. Includes data disk. An excellent companion for a more traditional introductory text. MK

Elementary Statistics, T(13). *Basic Statistics.* Stephen B. Jarrell. Wm C Brown, 1994, xix + 759 pp. [ISBN 0-697-21595-4] Classic introductory statistics text including a number of brief descriptions and analyses from published studies. Applications cover broad spectrum of scientific areas. Includes an enormous number of exercises. MK

Elementary Statistics, T(13: 1). *Statistics: A First Course, Instructor's Edition, Sixth Edition.* John E. Freund, Gary A. Simon. Prentice Hall, 1995, x + 597 pp. [ISBN 0-13-149949-1] Classical precalculus text with over half devoted to probability and descriptive statistics. Has chapters on the analysis of count data and nonparametric statistics, but only brief sections on analysis of variance and regression analysis. Emphasizes p-values. Gives answers to all exercises and complete solutions to many. (*Fourth Edition*, TR, May 1986.) RSK

Elementary Statistics, T(13: 1), C. *Elementary Statistics, Seventh Edition.* Robert Johnson. Duxbury Pr, 1996, xxvi + 836 pp, $61.25, with disk. [ISBN 0-534-24324-X] Most thorough revision of all editions. Presentation more visual, approachable, and clearer. Exercises improved and expanded with more real data (on disk). Simulation exercises, more focus on interpreting computer output, MINITAB instructions, and chapter case studies. Answers to selected exercises. (*Fourth Edition*, TR, December 1984). KB

Mathematical Statistics, T(15–17: 2). *A Course in Probability and Statistics.* Charles J. Stone. Duxbury Pr, 1996, ix + 838 pp, $80.75. [ISBN 0-534-23328-7] Assumes two years of calculus and some linear algebra. Probability portion makes more use of vectors and matrices than others at this level. Statistics portion of book has two major topics: normal models (and corresponding linear models); binomial and Poisson models and corresponding generalized linear models. Includes problems/answers. KB

Mathematical Statistics, T(15: 1). *Statistical Tests: An Introduction with MINITAB Commentary.* G.P. Beaumont, J.D. Knowles. Prentice Hall, 1996, ix + 285 pp, (P). [ISBN 0-13-842576-0] Nonstandard coverage, assuming a background of probability theory, including t- and F-distributions. Reviews basic MINITAB commands in Chapter 1. Nearly half the remaining text covers distribution free methods, with essentially nothing on analysis of variance and regression analysis. RSK

Mathematical Statistics, T*(16–17: 2, 3), L. *Introduction to Mathematical Statistics, Fifth Edition.* Robert V. Hogg, Allen T. Craig. Prentice Hall, 1995, xi + 564 pp. [ISBN 0-02-355722-2] Revision of the 1978 *Fourth Edition* (TR, December 1978), including some new material, additional exercises, and much reordering of the statistical topics. Sufficiency is now earlier, and nonparametrics is at the end. A solid text emphasizing the mathematics of mathematical statistics. RSK

Mathematical Statistics, T(14–15). *Informed Assessments: An Introduction to Information, Entropy and Statistics.* Alan Jessop. Ellis Horwood, 1995, xiii + 366 pp, (P). [ISBN 0-13-109229-5] Beautifully written introduction to information and uncertainty from both theoretical and decision-theoretic viewpoints. Exposition is excellent. An abundance of examples and exercises leads to facility with concepts such as "objective" vs. "subjective" probability, entropy, decision theory, and inference from joint distributions. MK

Statistical Methods, T(16–17), C. *Methods of Multivariate Analysis.* Alvin C. Rencher. Ser. in Prob. & Math. Stat. Wiley, 1995, xvi + 627 pp, $39.95, with disk. [ISBN 0-471-57152-0] Excellent introduction to standard topics in multivariate analysis. Plenty of exercises. Disk has data and SAS programs/commands for all exercises and examples. MK

Statistical Methods, T(16–17: 1, 2). *Multivariate Geostatistics: An Introduction with Applications.* Hans Wackernagel. Springer-Verlag, 1995, xiv + 256 pp, $59. [ISBN 3-540-60127-9] Statistical techniques for analysis of spatially- or temporally-correlated data.

Assumes knowledge of elementary statistics through multiple linear regression. LB

Statistics, S(15–18), P, L. *The Cambridge Dictionary of Statistics in the Medical Sciences.* B.S. Everitt. Cambridge Univ Pr, 1995, 274 pp, $19.95 (P); $44.95. [ISBN 0-521-47928-2; 0-521-47382-9] Definitions of approximately 2000 terms. Most terms are statistical, but also includes some relevant mathematical, computing, and genetic terms. General statistical terms have shorter entries than those relevant to medical science. Some definitions contain formulae or graphical material. Aimed for less mathematically sophisticated reader. Good reference for those in medical-related fields and students in applied statistics. KB

Statistics, T(14–16: 2).** *Statistics: Theory and Methods, Second Edition.* Donald A. Berry, Bernard W. Lindgren. Duxbury Pr, 1996, xi + 702 pp, $70. [ISBN 0-534-50479-5] Calculus-based course in statistical theory and methods. Notable for its emphasis on current statistical practices (including Bayesian approach where feasible), and liberal use of motivating examples from professional literature and industry. Nonparametric methods incorporated in sections on inference. Emphasizes distinction between use of data as evidence (in scientific studies), and for guiding decisions (in business and industry). (*First Edition*, TR, August–September 1990.) LB

Mathematical Computing, C, P. *Mathematica CD–ROM Library.* J. Braselton, *et al.* $55. [ISBN 0-12-059757-8] Five books on a CD: *Mathematica by Example, Revised Edition* (TR, May 1992), *The Mathematica Handbook* (TR, August–September 1993), and *Differential Equations with Mathematica* (TR, May 1994) by M. Abell and J. Braselton; *Mastering Mathematica* by J. Gray (TR, December 1994); *Introduction to Computer Performance Analysis with Mathematica* by A. Allen (TR, May 1994).

Computer Systems, C, P. *Compact Guide to Windows 95.* Jim Turley. Academic Pr, 1996, viii + 370 pp, $59.95 (P), with CD–ROM. [ISBN 0-12-703865-5] CD provides interactive introduction to Windows 95; book provides more details and covers additional topics.

Computer Systems, C, P. *Virus: Detection and Elimination.* Rune Skardhamar. Academic Pr, 1996, xiii + 290 pp, $34.95 (P), with disk. [ISBN 0-12-647690-X] Non-technical discussion of computer viruses and advice on how to detect and exterminate them. Code examples in 8086 assembly language. DOS disk contains disinfectant software.

Computer Systems, C, P. *Compact Guide to Visual Basic 4.* Bill Murray, Chris Pappas. Academic Pr, 1996, xiv + 417 pp, $49.95 (P), with CD–ROM. [ISBN 0-12-511910-0] CD provides interactive introduction to Visual Basic 4; book treats topics in more depth.

Computer Systems, C, P. *Introduction to Graphics Programming for Windows 95: Vector Graphics Using C++.* Michael J. Young. Academic Pr, 1996, xiii + 406 pp, $39.95 (P), with disk. [ISBN 0-12-773351-5]

Applications (Engineering), T(16). *Introduction to Random Processes in Engineering.* A.V. Balakrishnan. Wiley, 1995, xiii + 402 pp, $54.95. [ISBN 0-471-12487-7] Develops theoretical framework for processing of random signals and data. Emphasizes multidimensional analysis of linear time invariant systems, particularly "white noise." Wonderful review chapter dealing with linear algebra, probability, and Fourier series. Probability required for complete understanding as sample paths are utilized. In-depth treatment of linear filtering theory. Clear, well-written. Nice mathematical development. KS

Applications (Engineering), T(17–18), P. *Stability and Optimization of Flexible Space Structures.* S.J. Britvec. Birkhäuser Boston, 1995, xxvii + 280 pp, $134. [ISBN 0-8176-2864-9]

Applications (Mechanics), T(14–16), L. *The Sheer Joy of Celestial Mechanics.* Nathaniel Grossman. Birkhäuser Boston, 1996, xvii + 181 pp, $34.50. [ISBN 0-8176-3832-6] Accessible to any student who has completed several variable calculus. Author has an enthusiastic but breathless style, leaving many results to the exercises. Includes expansions for computing elliptic orbits, gravitational properties of solid bodies, shape of self-gravitating fluid. DB

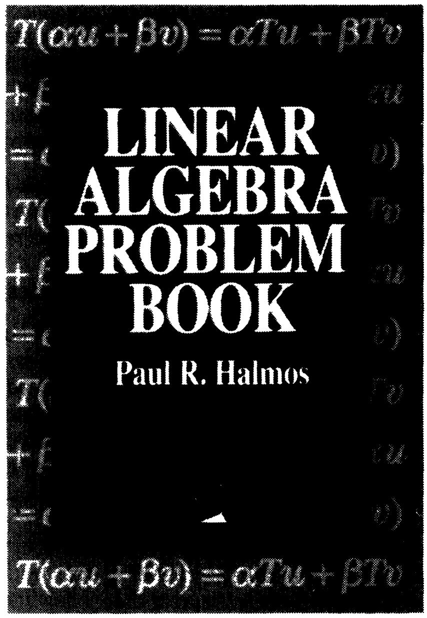
Applications (Physics), P. *Mach's Principle: From Newton's Bucket to Quantum Gravity.* Eds: Julian B. Barbour, Herbert Pfister. Einstein Stud., V. 6. Birkhäuser Boston, 1995, viii + 536 pp, $64.50. [ISBN 0-8176-3823-7] Papers and discussions from a 1993 conference held in Tübingen, as well as English translations of some important historic papers.

Reviewers

KB: Karla Ballman, Macalester; LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; JNC: Judith N. Cederberg, St. Olaf; LC: Laura Chihara, St. Olaf; JD-B: John Dyer-Bennet, Carleton; MK: Michael Kahn, St. Olaf; RSK: Richard S. Kleber, St. Olaf; RM: Richard Molnar, Macalester; TR: Tommy Ratliff, St. Olaf; JS: John Schue, Macalester; KS: Karla Spriestersbach, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf.

Linear Algebra Problem Book

Paul R. Halmos



Were it possible for the experience of apprenticeship to a master of mathematics to be packaged between the covers of a book, this would be it. No teacher of linear algebra should neglect to consult it. Highly recommended for all libraries.
– Choice Magazine

This is a book for mathematicians at all levels. Paul Halmos tells us, "Even if I know some answers, I don't think I understand a subject until I know the questions. The questions in mathematics are called problems—and although I learned some linear algebra a long time ago, until now I have made no serious effort to examine the problems that the solutions are based on. I wrote this book to organize those questions—problems—in my own mind."

This book is useful to anyone who needs linear algebra—and nowadays that means every user of mathematics. It can be used as the basis of either an official course or a program of private study.

If used as a course, the book can stand by itself, or if so desired, it can be stirred in with a standard linear algebra course as the seasoning that provides the interest, the challenge, the motivation that is needed by experienced scholars as much as by beginning students.

Contents

340 pp., Paperbound, 1995
ISBN 0-88385-322-1
List: $39.95
MAA Member: $28.95
Catalog Code: DOL-16/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	DOL-16/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires __/__

Signature ______________________

CRYPTOLOGY

Albrecht Beutelspacher

This fascinating little book is eminently readable, and it is a great deal of fun to peruse... the book is a real treat. We need more books like this, crafted by expert hands yet crafted so that the general reader can enjoy them.
—Bulletin of The Institute of Combinatorics and Its Applications

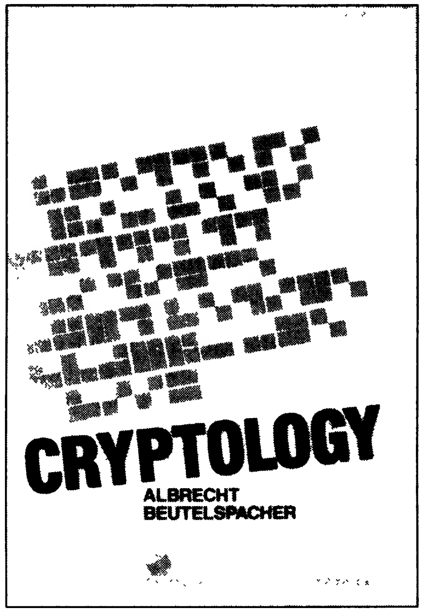
This excellent and entertaining book is suitable for a first course in cryptology for mathematical enthusiasts. An abundance of exercises and an excellent list of related references are included.
—The Mathematics Teacher

In spite of the light-hearted style in which the book is written throughout, it is a serious—and successful—attempt to explain the basis of coding and decoding messages...I can strongly recommend this book to anyone who wants a brief but comprehensive, eminently readable, and up-to-date introduction to this increasingly popular topic.
— The Mathematical Gazette

All of cryptology is covered in this work...Occupying a niche in the halls of the ivory tower of pure mathematics for nearly two millennia, number theory now forms a pillar of modern society. This book is the best explanation available today of how that pillar was constructed.
— Charles Aschbacher

A model to follow in order to make mathematics better known and understood. Accessible to a broad audience. Have fun reading this book, while you are getting a better understanding of cryptology.
— Bulletin of the Belgian Mathematics Society

How can messages be transmitted secretly? How can one guarantee that the message arrives safely in the right hands exactly as it was transmitted? Cryptology—the art and science of "secret writing"—provides ideal methods to solve these problems of data security.



The book is fun to read, and the author presents the material clearly and simply. Many exercises and references accompany each chapter.

176 pp., Paperbound, 1994
ISBN 0-88385-504-6
List: $34.00
MAA Member: $26.95
Catalog Code: CRYPT/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

Qty.	Catalog Code	Price	Amount
______	CRYPT/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

SPRINGER FOR MATHEMATICS

EDWARD LOZANSKY, National Science Teacher's Association, Washington DC and **CECIL ROUSSEAU**, University of Memphis, TN

WINNING SOLUTIONS

This book is intended to provide students with the appropriate mathematical tools and problem-solving experience to successfully compete in high-level problem solving competitions. In each section, the authors attempt to "fill in" the appropriate background and then provide the student with a variety of worked examples and exercises to help bridge the gap between what he or she may already know and what is required for high-level competitions. Answers or sketches of the solutions are given for all exercises. The book makes an attempt to introduce each area "gently" assuming little in the way of prior background — and teach the appropriate techniques, rather than simply providing a compilation of high-level problems.

1996/APP. 267 PP./SOFTCOVER/$45 00 (TENT)
ISBN 0-387-94743-4
PROBLEM BOOKS IN MATHEMATICS

DARREN REDFERN, Practical Approach Corp, Waterloo, Canada and **EDGAR CHANDLER**, Paradise Valley Community College

MAPLE ODE LAB BOOK

This is the first lab book that utilizes the computational power of Maple V to efficiently illustrate the solving of ordinary differential equations. This package will provide students with a variety of methods and an environment where they can experiment and develop new ways of tackling difficult concepts. The manual begins with a well-rounded introduction to Maple as a symbolic computational language so that a background in Maple is not necessary.

1996/APP 150 PP, 50 ILLUS, 3.5" DISK
HARDCOVER/$29 95 (TENT)
ISBN 0-387-94733-7

Second Edition

JOSEPH BAK, City College, CUNY, NY and **DONALD J. NEWMAN**, Temple University, Philadelphia, PA

COMPLEX ANALYSIS

This unusually lively textbook on complex variables introduces the theory of analytic functions, explores its diverse applications and shows the reader how to harness its powerful techniques. **Complex Analysis** offers new and interesting motivations for classical results and introduces related topics that do not appear in this form in other texts. Stressing motivation and technique, and containing a large number of problems and their solutions, this volume may be used as a text both in classrooms and for self-study. Topics covered include: the complex numbers; functions of a complex variable; analytic functions; line integrals and entire functions; properties of entire functions and of analytic functions; simply connected domains; isolated singularities; the residue theorem and applications; contour integral techniques; conformal mapping and the Riemann mapping theorem; maximum-modulus theorems for unbounded domains; harmonic functions; forms of analytic functions; analytic continuation; the gamma and zeta functions; application to other areas of mathematics. For this second edition, the authors have revised some of the existing material and have provided new exercises and solutions.

1996/APP 253 PP, 69 ILLUS./HARDCOVER
$39.95/ISBN 0-387-94756-6
UNDERGRADUATE TEXTS IN MATHEMATICS

Forthcoming...

CHUANMING ZONG, Academia Sinica, Beijing, China and **JAMES J. DUDZIAK**, Bucknell College, PA

STRANGE PHENOMENA IN CONVEX AND DISCRETE GEOMETRY

1996/APP 192 PP /SOFTCOVER/$ 29 00
ISBN 0-387-94734-5
UNIVERSITEXT

Springer

HORST STÖCKER, Johann Wolfgang Goethe Universität, Frankfurt, Germany

THE HANDBOOK OF MATHEMATICAL FORMULAS

This is a comprehensive handbook of mathematical formulas and theorems which will be invaluable to students, mathematicians and scientists who are in need of exact statements of mathematical results. This indispensable volume will become the ultimate desktop reference for all mathematicians and students.

1996/APP. 952 PP., 545 ILLUS /HARDCOVER
$29 95/ISBN 0387-94746-9

LARRY GERSTEIN, University of California, Santa Barbara

INTRODUCTION TO MATHEMATICAL STRUCTURES AND PROOFS

1996/350 PP, 103 ILLUS /HARDCOVER/$52.95
ISBN 0-387-97997-2
TEXTBOOKS IN MATHEMATICAL SCIENCES

KATHLEEN ALLIGOOD, TIM SAUER, both of, George Mason University and **JAMES A. YORKE**, University of Maryland

CHAOS

An Introduction to Dynamic Systems

1996/APP 400 PP, 200 ILLUS /HARDCOVER
$25 00 (TENT.)/ISBN 0-387-94677-2
TEXTBOOKS IN MATHEMATICAL SCIENCES

Instructors: For an exam copy of the Gerstein or Alligood textbooks, please contact Jones & Bartlett Publishers at 1-800-832-0034 or e-mail springer@jbpub com

Visit the Springer Website...
Springer America (http //www springer-ny com)
Springer Europe (http://www springer de)
Telos (http //www telospub com)
Subscribe to our Listserver (send an e-mail with the word *HELP* in the body of the message to listproc@lists springer-ny.com)

Order Today!

- **Call:** 1-800-SPRINGER or **Fax:** 201-348-4505 (8 30 am-5.30 pm ET)
- **Write:** Send payment (check or credit card) plus $3 00 postage and handling for the first book and $1 00 for each additional book to Springer-Verlag New York, Inc, Dept #S237, PO Box 2485, Secaucus, NJ 07096-2485 (CA, IL, MA, NJ, NY, PA, TX, VA, and VT residents add sales tax, Canadian residents add 7% GST)
- **Visit:** Your local technical bookstore
- **E-mail:** orders@springer-ny.com
- **Instructors:** Call or write for info on textbook exam copies

6/96 Reference S237

THE AMERICAN MATHEMATICAL MONTHLY

Volume 103, Number 7 — August–September 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Catalan's Conjecture

Paulo Ribenboim

Catalan's conjecture is very simple to state. The aim of this lecture is two-fold:

1°) To describe the methods used in trying to solve the problem.
2°) To explain why the problem is not only challenging, but also natural.

If you find it also beautiful, I have succeeded. With one exception, proofs are omitted but ample references are provided.

1. THE PROBLEM. I shall consider sequences of integers and ask some questions. First, consider the sequence of all squares or cubes:

$$4, 8, 9, 16, 25, 27, 36, 49, 64, 81, 100, \ldots$$

It may be observed that 8 and 9 are consecutive numbers in this sequence. The first problem is:

Are there any other consecutive integers in the above sequence? How many pairs of consecutive integers? Finitely many? Infinitely many?

I may also consider the sequence of *all* proper powers, which includes also 5th powers, 7th powers, 11th powers, etc... (note that powers with even exponents are squares, powers with exponent multiples of 3 are cubes...)

The same question may be asked: Are there consecutive powers other than 8 and 9?

But for the sequence of all powers, a new problem makes sense: Are there three consecutive integers that are proper powers?

Since powers grow very fast, lists of powers are necessarily very limited and, besides 8 and 9, no consecutive powers have ever been observed. This is an indication to keep in mind, but one should be careful before jumping to any conclusion.

Just think, for example, that up to 100, 10% of the numbers are squares, up to 10,000, 1% are squares, up to 1,000,000, 1 in 1,000 are squares, and so on. Yet, Lagrange proved that despite the increasing scarcity of squares, every natural number is the sum of at most 4 squares. The squares seem to occupy strategic positions. Of course, ours is a different problem.

Similar problems may be asked with the following sequence. Let a, b be integers, $1 < a < b$ and consider the sequence of all powers of a or of b. For example, if $a = 2$, $b = 3$, it is the sequence

$$4, 8, 9, 16, 27, 32, 64, 81, \ldots$$

How many pairs of consecutive integers may be found in such sequences?

Further, let E be a finite (nonempty) set of prime numbers, and let $E^{\times}$ be the set of all natural numbers whose prime factors belong to E. How many pairs of consecutive integers belong to $E^{\times}$?

All the preceding problems may be easily expressed in terms of diophantine equations.

The first problem amounts to the solution in natural numbers of the equations

$$X^2 - Y^3 = 1, \qquad X^3 - Y^2 = 1.$$

The problem concerning arbitrary powers is expressed by the exponential diophantine equation, using unknowns

$$X^U - Y^V = 1$$

to be solved in integers greater than 1.

If $1 < a < b$, the third problem is the same as the solution in integers greater than 1, of the equations

$$a^U - b^V = 1, \qquad b^V - a^U = 1.$$

Finally, the problem for the sequence $E^{\times}$, corresponds to the simple equation

$$X - Y = 1,$$

but the solutions have to belong to $E^{\times}$.

In 1844, Catalan conjectured that 8, 9 are the only consecutive integers that are powers.

Despite much progress—which I'll soon describe—Catalan's conjecture has yet to be proved or disproved.

2. RELATION WITH OTHER PROBLEMS. It is my purpose now to convince you that the problems indicated belong to a natural and broad class of interesting and well-known questions. So, you will not find Catalan's problem irrelevant to a better understanding of the integers.

Let P be a set of natural numbers; whenever convenient, it may be assumed that $0 \in P$.

I shall describe addition and subtraction problems.

Addition Problems. Let $P + P = \{p + p' | p, p' \in P\}$. If $n \geq 1$, let $nP = \{p_1 + p_2 + \cdots + p_n | \text{each } p_i \in P\}$.

Let $\langle P \rangle = \bigcup_{n \geq 1} nP$.

One wishes to study the sets nP, $\langle P \rangle$ and compare them with the set $\mathbb{N}$ of all natural numbers or with some appropriate subset of $\mathbb{N}$.

For example, these are the usual questions:

Does there exist n such that $nP = \mathbb{N}$? Is $\langle P \rangle = \mathbb{N}$?

There are also the corresponding asymptotic questions. Does there exist k_0 such that

$$\{k \in \mathbb{N} | k \geq k_0\} \subseteq nP \quad \text{or} \quad \{k \in \mathbb{N} | k \geq k_0\} \subseteq \langle P \rangle?$$

In such situations, can one find k_0 effectively?

Subtraction Problems. Now the problem is to identify the set $P - P = \{p_1 - p_2 | p_1, p_2 \in P\}$.

More precisely, if $n \in P - P$, we wish to determine the set $\{(p, p') \in P \times P | n = p - p'\}$ or at least find bounds for the number of elements of the set.

Again, in some cases, the answer is known only asymptotically and it may be quite difficult.

These ideas will now be illustrated in some specific situations.

1) Prime numbers. Let P be the set of all prime numbers. More generally, if $k \geq 1$ let P_k be the set of all integers of the form $p_1^{e_1} \cdots p_n^{e_n}$ with $0 < e_1 + \cdots + e_n \leq k$. The integers in P_k are called k-almost primes.

Thus $P_1 = P$.

Addition problem: Goldbach problem.

The famous conjecture of Goldbach states that

$$\{2n | n \geq 2\} \subset P + P,$$

or equivalently,

$$\{n | n \geq 6\} = P + P + P.$$

In my book on prime numbers (cited in the references), I described the main results obtained in the study of Goldbach's conjecture.

For example, Vinogradov proved:

$$\{n | n \text{ odd}, n \geq 3^{3^{15}}\} \subset P + P + P.$$

Schnirelmann showed that there exists S_0 such that

$$\{n | n \geq 2\} = \bigcup_{k=1}^{S_0} kP.$$

Riesel and Vaughan calculated that S_0 may be taken to be 19.

Allowing almost primes, I note the pioneering result of Brun:

$$\{n | n \geq 4\} = P_9 + P_9.$$

The best result known today is due to Chen:

$$\{n | n \geq 4\} = P + P_2.$$

Subtraction problems: Polignac's conjecture and twin primes conjecture.

Polignac conjectured that every even number is the difference of two primes; in other words:

$$\{2k | k \geq 1\} \cup \{1\} = P - P.$$

This conjecture has never been proved or disproved.

The twin primes conjecture is the statement that there exist infinitely many primes p such that $p + 2$ is also a prime. In other words, 2 may be represented in infinitely many ways in the form $2 = p' - p$, where p, p' are primes. This statement is also waiting for a proof.

For each $N > 1$, let $\pi_2(N)$ denote the number of primes $p \leq N$ such that $p + 2$ is also prime. The following is a quantitive version of the twin primes conjecture:

$$\pi_2(N) \sim \frac{N}{(\log N)^2},$$

that is, the quotient of the two expressions has a limit equal to 1 (as $N \to \infty$).

According to Brun, twin primes are scarce, since

$$\sum \frac{1}{p} = B < \infty$$

(sum over all primes p such that $p + 2$ is also a prime). Note that $\Sigma(1/p) = \infty$ (sum for all primes), as was proved by Euler.

2) Powers and powerful numbers. Let P be the set of all proper powers. Let Q be the set of all powerful numbers (that is, numbers N such that if p divides N, then p^2 divides N).

It is immediate that $Q = \{a^2 b^3 | a, b \geq 1\}$.

Addition problems. An interesting problem concerning the set $P + P$ is the description of $(P + P) \cap P$; in other words, the study of the solutions of $X^l + Y^m = Z^n$ for fixed l, m, n or even arbitrary l, m, n. In particular, the study of the equation $X^n + Y^n = Z^n$ (Fermat's equation) has extended over three centuries. The problem of Fermat has just been solved by A. Wiles (with the collaboration of R. Taylor), who proved:

If $n \geq 3$ and x, y, z are natural numbers such that $x^n + x^n = z^n$, then $xyz = 0$.

The situation is very different when $n = 2$. It has long been known that there exist infinitely many triples of pairwise relatively prime integers (x, y, z) such that $x^2 + y^2 = z^2$ (these are the Pythagorean triples).

A similar result has been recently obtained by Elkres: there exist infinitely many fourth powers that are sums of three fourth powers.

Another famous addition problem is due to Waring. Given $k \geq 2$, does there exist an integer $G(k) > 1$ such that every sufficiently large natural number is the sum of at most $G(k)$ kth powers? Similarly, does there exist an integer $g(k) > 1$ such that every natural number is the sum of at most $g(k)$ kth powers?

In this respect—as I have already mentioned—Lagrange proved that for squares, $g(2) = 4$, while Gauss showed that $G(2) = 4$ because there are infinitely many rational numbers that are not the sum of 3 squares.

Hilbert showed the existence of $g(k)$ for each $k \geq 2$. The problem became the exact calculation of $G(k)$ and $g(k)$. Thus, Davenport showed that $g(4) = 19$. The complete solution for 4th powers was given recently by Balasubramanian, Deshouillers, and Dress: $G(4) = 16$. Explicitly, all sufficiently large integers are sums of 16 fourth powers; there exist infinitely many integers that are not sums of 15 fourth powers; all integers are sums of at most 19 fourth powers.

More results about Waring's problem are gathered in my book on prime numbers.

Concerning powerful numbers, I note that not every natural number is the sum of two powerful numbers. On the contrary

$$\lim_{N \to \infty} \frac{\#\{n \in Q + Q | n \leq N\}}{N} = 0.$$

However Heath-Brown has shown that every sufficiently large natural number is the sum of at most three powerful numbers.

Subtraction problems. This time I consider first the powerful numbers. The notation $1 \underset{\infty}{\in} Q - Q$ means that 1 is in infinitely many ways the difference of powerful numbers; in other words, there exist infinitely many pairs of consecutive powerful numbers. Indeed, there are infinitely many pairs (x, y) such that $x^2 - 8y^2 = 1$; thus $x^2, 8y^2$ are consecutive powerful numbers.

With similar notation, Mollin and McDaniel showed that $n \underset{\infty}{\in} Q - Q$, for every $n \geq 2$.

Concerning three consecutive powerful numbers, Erdös conjectured: there do not exist three consecutive powerful numbers.

Granville showed how to deduce from this conjecture the theorem of Adleman, Heath-Brown, and Fouvry: there exist infinitely many primes p such that if x, y, z are natural numbers and $x^p + y^p = z^p$, then p divides xyz (first case of Fermat's last theorem). Despite the recent proof of Fermat's last theorem, the connection between this theorem and powerful numbers remains intriguing.

The corresponding question for powers amounts to Catalan's conjecture: if $1 = p' - p$ (with $p, p' \in P$) then $p' = 9$, $p = 8$.

Pillai conjectured: for every $k > 1$ there exist only finitely many pairs of powers (p, p') with $p, p' \in P$ and $k = p' - p$.

Pillai's conjecture may be expressed in terms of the sequence

$$z_1 < z_2 < z_3 < \cdots$$

of all powers, namely,

$$\lim_{i \to \infty} (z_{i+1} - z_i) = \infty.$$

At the appropriate moment, I shall deal with three consecutive powers.

INTERLUDE. Having explained the significance of Catalan's conjecture, in relation to other well-known and important problems, I feel that I may count on your interest. My task will be to describe the numerous attempts to solve the problem.

If you have a question to solve and you cannot do it right away, it is wise to assess its difficulty by looking at special cases. A careful analysis of successes in special situations, leads to more systematic and embracing methods; for Catalan's problem it involves algebraic numbers, and in the last analysis, congruences. However, by their own nature these beautiful methods cannot cover all the possibilities; this will become apparent later.

So, what to do? Guided by the intuitive feeling of more and more rarified powers, analytical methods should discern whether there are solutions with arbitrarily large numbers. Be patient until I explain how the theory of diophantine approximation has provided the most successful method to attack the problem.

3. SPECIAL CASES. The first recorded result in connection with the problems of Catalan and analogues dates back to around 1320 and is due to Levi ben Gerson (= Leo Hebraeus), a famous astronomer of his time. He proved that if powers of 2 and 3 are consecutive, then they must be 8 and 9. Today this is an easy exercise with congruences.

Euler proved that if $X^2 - Y^3 = \pm 1$ then $X = 3$ and $Y = 2$. Here is the idea behind the proof that $X^2 - Y^3 = -1$ has no solution in integers $x, y > 0$. Indeed, if $x^2 - y^3 = -1$ then $y^3 = x^2 + 1 = (x + i)(x - i)$, where $i^2 = -1$. From the arithmetic of Gaussian integers (easy facts known to Euler), $x + i = \alpha(a + bi)^3$, where a, b are integers and $\alpha = \pm 1$ or $\pm i$. Then $x - i = \bar{\alpha}(a - bi)^3$ with $\bar{\alpha} = \pm 1$ or $\mp i$ (respectively). Then $2i = \alpha(a + bi)^3 - \bar{\alpha}(a - bi)^3$, and an easy calculation shows that this is impossible. The fact to note is the appeal to Gaussian integers. This idea, duly modified, is found also in the study of other special cases. This is embodied in the following lemma preceded by an obvious remark. If $m, n \geq 2$ and $x^m - y^n = 1$, let p, q be primes, $m = pm'$, $n = qn'$, then $(x^{m'})^p - (y^{n'})^q = 1$. Thus, to show that $X^m - Y^n = 1$ has no solution it suffices to consider the same equation, when the exponents are primes p, q.

Now if p, q are odd primes, $x, y \neq 0$, and $x^p - y^q = 1$, then $y^q = x^p - 1 = (x - 1)\left(\frac{x^p - 1}{x - 1}\right)$.

Since $\gcd\left(x - 1, \frac{x^p - 1}{x - 1}\right) = 1$ or p, two cases are possible:

$$\begin{cases} x - 1 = r^q \\ \dfrac{x^p - 1}{x - 1} = r'^q \end{cases}$$

with $\gcd(r, r') = 1$ and $rr' = y$, or

$$\begin{cases} x - 1 = p^{q-1} r^q \\ \dfrac{x^p - 1}{x} = pr'^q \end{cases}$$

with $\gcd(r, r') = 1$ and $prr' = y$ (since p^2 does not divide $\frac{x^p - 1}{x - 1}$).

From $x^p = y^q + 1 = (y + 1)\left(\frac{y^q + 1}{y + 1}\right)$, one obtains analogous expressions for $y + 1$, $\frac{y^q + 1}{y + 1}$ in two cases.

There are also similar expressions derived from $x^2 - y^q = 1$ (with q odd prime).

The next special cases dealt with $X^2 - Y^q = 1$, respectively $X^p - Y^2 = 1$ (with p, q primes greater than 3).

Now it happened that one of the preceding equations was analyzed without difficulty and was solved in 1850 by Lebesgue only six years after Catalan announced his conjecture (1844), whereas, the other equation, despite multiple attempts, required 120 years to be finally solved by Ko in 1964.

Which one is which?

This question is very "á-propos" in order to stress that sometimes two Diophantine equations may look very much alike but their solution demands methods of very different levels of difficulty.

Lebesgue proved, with a variant of the method of Euler, that $X^p - Y^2 = 1$ (with p prime $p \geq 5$) has only trivial solutions.

The proof of Ko (in 1964) that $X^2 - Y^q = 1$ (q prime, $q \geq 5$) has only trivial solutions, was much more contorted. Later, Chein used results of Stǿrmer and Nagell from the beginning of this century, to give a clever and much shorter proof of Ko's theorem. Only three pages sufficed!

The study of the equations $X^3 - Y^q = 1$, $X^p - Y^3 = 1$ (for p, q primes greater than 3), leads to the equations

$$X^2 + X + 1 = Y^q$$

or

$$X^2 + X + 1 = 3Y^q.$$

These equations were considered by Nagell, who stated that they have only trivial solutions, provided the solutions of the equation

$$X^3 - 3XY^2 + Y^3 = 1$$

were the ones already known: $(x, y) = (1, 0), (0, 1), (-1, -1), (2, -1), (1, 3)$ and $(-3, -2)$. This was not easy to establish; Ljunggren (1942) succeeded with a precise analysis of the group of units in a certain cubic field.

I like to stress that no one dared to attack the equation $X^p - Y^q = 1$ where $\min\{p, q\} \geq 5$, using ad-hoc special methods.

4. ALGEBRAIC METHODS. The purpose of the methods relying heavily on the arithmetic of algebraic number fields is to treat simultaneously large classes of exponents. Congruences, units, and classes of ideals abound in these considerations.

But first I wish to list some additional conditions that imply that the only non-trivial solution of $X^U - Y^V = 1$ (with exponents at least 2) is $x = 3$, $y = 2$, $u = 2$, $v = 3$, giving $9 - 8 = 1$.

Namely:

a) If p, q are primes, l prime and $l^p - y^q = \pm 1$, then necessarily $l = 3$, $p = 2$, $y = 2$, $q = 3$.

b) If $x, y \geq 2$ and $x^y - y^x = 1$, then $x = 3$, $y = 2$.

c) The only consecutive powers of consecutive integers are 8, 9, in other words, $x^m - y^n = 1$ and $|x - y| = 1$ imply that $x = 3$, $y = 2$, $m = 2$, $n = 3$.

The proof of c) requires an interesting long-known arithmetical result on prime divisors of expressions of the form $x^m - 1$.

Cassels gave a remarkable proof of the following result:

If $x^p - y^q = 1$ (with p, q primes), then p divides y and q divides x.

It follows that in the old Euler's lemma, only the second case can actually take place. Thus $x - 1 = p^{q-1}r^q$, $\frac{x^p - 1}{x - 1} = pr'^q$ and also $y + 1 = q^{p-1}s^p$, $\frac{y^q + 1}{y + 1} = qs'^p$.

One wonders what would be the importance of Cassels' result. Not knowing the existence of x, y such that $x^p - y^q = 1$, how can one use the fact that $p|y$ and $q|x$?

Surprise! Both Hyyrö (in Finnish) and Mąkowski proved:

There do not exist three consecutive powers.

It seems to be an unwritten rule that every lecture should include at least one proof. So I choose this one for its striking simplicity:

If $x^p < y^q < z^r$ are proper powers with exponents that may be taken to be primes, if $y^q - x^p = 1$, $z^r - y^q = 1$, then by Cassels' result $q|x$ and $q|z$. Hence $q|x^p$, $q|z^r$, so q divides their difference $z^r - x^p = 2$. Thus $q = 2$ and so $z^r - y^2 = 1$. But this is impossible by the result of Lebesgue. Contradiction and end of proof.

The theorem of Cassels implies that if $x^p - y^q = 1$, then x, y are of special form, namely $x = 1 + p^{q-1}r^q$, $y = -1 + q^{p-1}s^p$, and also $\frac{x^p - 1}{x - 1}, \frac{y^q + 1}{y + 1}$ are of special form.

Hyyrö explored this idea, giving more restrictions that must be satisfied by x, y. But above all, he followed the lead of Wieferich and Inkeri to relate the problem to the congruences obtained by Wieferich for Fermat's last theorem. I explain now the very useful results of Inkeri, which continued the line of Hyyrö's research.

Let p be an odd prime and denote by $H(-p)$ the class number of the field $\mathbb{Q}(\sqrt{-p})$. Here is one of Inkeri's results:

Let $p > 3$, $p \equiv 3 \pmod 4$. If q is a prime, $q > 3$ and

$$\begin{cases} q \nmid H(-p) & \text{and} \\ p^{q-1} \not\equiv 1 & \pmod{q^2} \end{cases}$$

then $X^p - Y^q = 1$ has only trivial solutions.

Inkeri gave a similar criterion when $q \equiv 3 \pmod 4$ and also a stronger criterion when both $p \equiv 3 \pmod 4$ and $q \equiv 3 \pmod 4$, all of this complemented with further precision in special cases.

The interest of these results for practical purpose is twofold. First, it is relatively easy to calculate the class number of an imaginary quadratic field and to check if a given prime divides it. Second, it has been observed that the so-called Wieferich congruence (with base p) $p^{q-1} \equiv 1 \pmod{q^2}$ occurs very rarely. This and similar

criteria allow us, after computation, to decide that for many pairs of exponents (p, q) the corresponding equation has only trivial solutions.

But even a small pair like $(5,7)$ cannot be dealt with by this criterion. Indeed, $q = 7 \equiv 3 \pmod 4$, $H(-7) = 1$, 5 does not divide $H(-7)$; however, $7^4 \equiv 1 \pmod{5^2}$.

To cover more cases, Inkeri considered cyclotomic fields also. Let h_p denote the class number of the cyclotomic field $\mathbb{Q}(\zeta_p)$ where ζ_p is a primitive pth root of 1.

Inkeri showed:

Assume that $X^p - Y^q = 1$ has a non-trivial solution.

1) If p does not divide h_q then $q^{p-1} \equiv 1 \pmod{p^2}$.
2) If q does not divide h_p, then $p^{q-1} \equiv 1 \pmod{q^2}$.

In particular, the equations $X^5 - Y^7 = \pm 1$ have only trivial solutions. Indeed $5 \nmid h_7$, $7 \nmid h_5$ but $5^6 \not\equiv 1 \pmod{7^2}$. In a subsequent paper with Aaltonen, many more pairs of exponents were disposed of by this method, after computation of class numbers and Wieferich congruences.

These calculations have been pushed up by Mignotte. The last word is that (with a still unpublished lemma by W. Schwarz) if $\min\{p, q\} \le 10{,}640$, then $X^p - Y^q = 1$ has only trivial solutions.

5. ANALYTICAL METHODS. At this moment, I would like to stress what is obvious and has been implicit. Namely, equations of three different types have been under consideration:

I. $a^U - b^V = 1$ where a, b are given distinct integers greater than 1.
II. $X^m - Y^n = 1$, where m, n are distinct integers greater than 1.
III. $X^U - Y^V = 1$.

So, it is appropriate to discuss in turn each one of these equations.

I. Equation $a^U - b^V = 1$. The main result is by LeVeque, who showed that there exists at most one pair (u, v), with $u \ge 2$, $v \ge 2$ such that $a^u - b^v = 1$.

Cassels gave an algorithm to find the hypothetical solution (if it exists). For $(a, b) \ne (3,2)$ the algorithm has—up to now—failed to find any solution!

I also want to consider the variant of this equation mentioned at the beginning. Let $E = \{p_1, \dots, p_s\}$ (with $s \ge 1$) be a finite set of primes. Let $k \ge 1$.

Thue proved that there exists an effectively computable constant $C > 0$ such that if

$$p_1^{n_1} p_2^{n_2} \cdots p_s^{n_s} \cdots p_1^{m_1} p_2^{m_2} \cdots p_s^{m_s} = k$$

(with integers $n_i, m_i \ge 0$), then $n_i, m_i < C$ (for all $i = 1, \dots, s$).

The special cases when $k = 1$ or 2 had been proved earlier by Størmer with a very interesting method involving divisibility properties of terms of linear recurring sequences of order 2 (in other words, analogues of the sequences of Fibonacci numbers and of Lucas numbers).

II. Equation $X^m - Y^n = 1$. Siegel dealt with a more general equation. From his main result, it follows:

If $m, n \ge 2$ with $\max\{m, n\} \ge 3$, and if a, b, k are given non-zero integers, then the equation $aX^m - bY^n = k$ has only finitely many solutions in integers.

The result of Siegel did not include any bound on the number or, a fortiori, on the size of the eventual solutions.

It was Baker's great achievement, which earned him the Fields Medal, to invent a new method leading to effective bounds on eventual solutions of many types of diophantine equations. In the present case, Baker's estimates gave:

If $m, n \geq 2$, $k \geq 1$, and $x^m - y^n = k$, then $|x|, |y| < \exp\exp((3m)^{10} n^{10n^3} |k|^{n^2})$ (and a similar bound exchanging m with n). The bound depends on the strength of estimates of lower bounds for certain linear forms in logarithms. It should be noted that this bound involves a double exponentiation and is therefore very, very large.

It should also be mentioned that for the number of pairs (m, n) such that $X^m - Y^n = 1$ has a non-trivial solution, Hyyr found the following upper bound: $\exp(631 m^2 n^2)$. Smaller than Baker's, but bigger than 0—the hoped-for bound!

A good support to the conjecture comes from the following density theorem, which I proved using a theorem of Schinzel & Tijdeman: Given a, b, k non-zero integers for each $N > 1$ consider the number $\alpha(N)$ of pairs (m, n) with $2 \leq m$, $n \leq N$ such that the equation $aX^m - bY^n = k$ has no solution in positive integers. Then $\frac{\alpha(N)}{N^2}$ has limit (as N tends to ∞) that is equal to 1.

III. Equation $X^U - Y^V = 1$. It is time to state the most significant result thus far obtained about Catalan's conjecture. It was proved in 1976 by Tijdeman, who used Baker's inequalities twice, in a novel, clever way:

There exists a constant C such that if p, q are primes, x, y are positive integers, and $x^p - y^q = 1$, then $p, q < C$.

Coupled with the effective result of Baker for the equation (II), one may state:

There is a constant $T > 0$ such that if $x^p - y^q = 1$ with p, q primes and $x, y \geq 1$, then $x, y, p, q < T$.

Langevin estimated that T may be taken to be $\exp\exp\exp\exp(730)$—a number of size defying my imagination (just to think about it, I get a headache).

This theorem does not yet establish the truth of Catalan's conjecture. But it shows that the problem of Catalan is decidable in finitely many steps. Theoretically (if not in practice), it suffices to try, one after the other, all quadruples (x, y, p, q) up to the bound T, and to check if $x^p - y^q = 1$.

The consideration of sharper forms of Baker's inequalities in close connection with Catalan's equation has led Mignotte, on the one hand, and Glass (and his collaborators), on the other, into a race to lower the bound for the exponents. It is known that if $X^p - Y^q = 1$ has a non-trivial solution, then $\max\{p, q\} \leq 10^{26}$.

So, we know that Catalan's conjecture is decidable, but it is not known when it will be decided.

6. CONCLUSION. Once again, an innocent-looking problem of natural numbers turns out to be a challenge even for the most astute mathematicians. A careful study of the efforts to solve the problem might be compared, as I said in the preface to my book, to a journey through a beautiful mathematical landscape. A winding road with bright flowers to pick. A distant peak, which seems no more out of reach.

REFERENCES

For proofs, remarks, details concerning Catalan's conjecture, you may consult my own book, which contains a fairly complete bibliography about the problem:

P. Ribenboim, *Catalan's Conjecture*, Academic Press, Boston, 1994.

Previously, I have published a survey on the problem:

P. Ribenboim, Consecutive powers, *Expositiones Mathematicae* 2 (1984), 193–221.

Several very recent preprints by A. W. Glass et al., by M. Migotte, and by W. Schwarz are concerned with developments in the line of K. Inkeri's criteria and the related calculations, some of which are still unpublished.

The results about prime numbers may be found in my own book:

P. Ribenboim, *The Book of Prime Number Records*, Springer-Verlag, New York (first edition 1987; second edition 1989; third edition 1995).

See also the French abridged version:

P. Ribenboim, *Les Nombres Premiers: Mystères et Records*, Presses Universitaires de France, Paris, 1994.

Department of Mathematics and Statistics
Queen's University
Kingston, Ontario
K7L 3N6 CANADA
mastdept@qucdn.queensu.ca

Identify the following sequence: una, dues, cinc, catorze, quaranta-dues, cent trenta-dues, quatre-cent vint-i-nou, … .

(See page 577)

Real Cauchy sequences converge

A real Cauchy sequence (x_n) is bounded and so lies in some interval I_1 of length L. Trisect I_1 to obtain three subintervals of length $L/3$. As one of the two outer subintervals contains only finitely many x_n, a subinterval I_2 of I_1 of length $2L/3$ contains all but finitely many x_n. Repeating this gives I_n with $\bigcap I_n = \{y\}$ and $x_n \to y$.

Contributed by A. F. Beardon, University of Cambridge

Mathematics and Modern Technology

Roger S. Pinkham

This paper is an attempt to illustrate and emphasize three points. First, that modern technology allows one to encourage students to ask and answer questions heretofore impossible to address fruitfully. Second, many (probably most) classical methods are as necessary as ever, but perhaps in a different setting, and third, simple calculus and a bit of reflection is amazingly effective. Although the examples presented are specific, I have attempted to approach each in a manner that has general applicability. Throughout I have tried to show how modern technology can provide insight and foster a spirit of inquiry.

Machines do some things very well, some poorly. The same is true of humans. The two are often complementary; it seems best to search for uses of each that capitalize on their individual strengths. The topic of infinite series provides convenient examples. Given an infinite series, one asks two things. Does it converge, and if so, to what? The former question has been the concern of most calculus texts, while the latter has usually been left in abeyance. This is unfortunate for the applications-oriented student, for it is the value of the series that is most often the thing of primary concern.

Sums and integrals are intimately related. Quadrature formulae of the form

$$\text{Integral} = \text{Sum} + \text{Remainder}$$

can often be expeditiously employed to evaluate a sum provided the integration can be carried out and the remainder suitably bounded. Alternatively, one may alter individual terms in a sum slightly in hopes of obtaining upper and lower bounds for the sum in question. But having done so, it may still be necessary to use a machine to provide actual numerical estimates. Often one sums a few terms on the front end of a series and then deals with the resulting series for the remainder as if this were the original series. Another common trick is to find a second series with known sum and with asymptotic properties similar to those of the first. This second series is then subtracted from the first to increase the rate of convergence. This new series of differences can then be attacked with the previously-mentioned tools. One standard approach I shall not use is the Euler-Maclaurin summation formula.

What follows are intended as illustrative examples.

A KEPLERESQUE CONSTRUCTION. If you look at Figure 3 on page 224 of the classic *The World of Mathematics* by J. R. Newman [3], you will see inscribed and circumscribed spheres, representing the five regular solids as Kepler supposed them to be among the planetary orbits. In the late 50's my dear friend R. W. Hamming fashioned a related two-dimensional construction. He was visiting Dartmouth to examine the prospects of the newly constructed Kewit computer network. As an example of the kind of problem appropriate for such a terminal-based system, he considered the following geometric construction.

A unit circle is circumscribed with an equilateral triangle; the triangle is circumscribed by a circle. This circle is circumscribed with a square, this with a circle, thence a regular pentagon, etc. Figure 1 illustrates this process through $n = 15$.

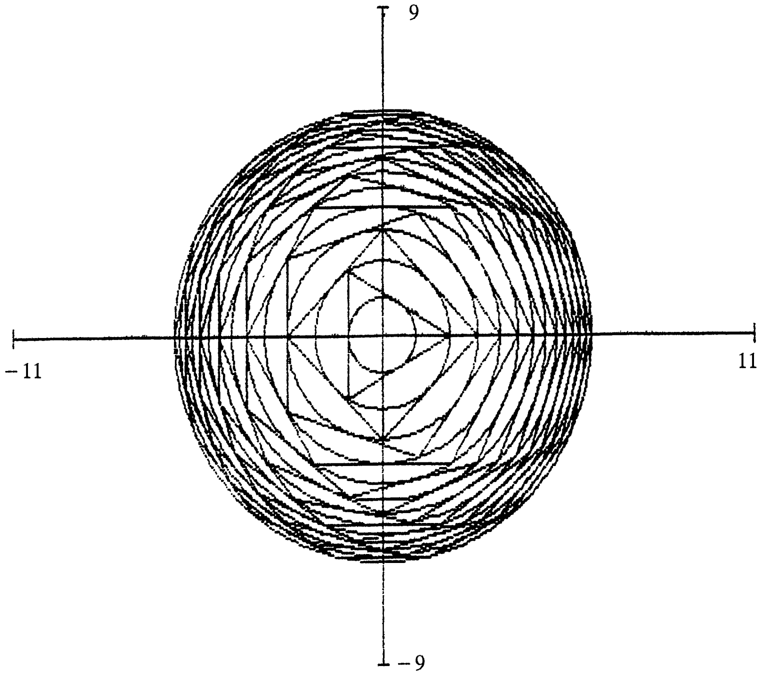


Figure 1. $n = 15$

Table 1 gives the radii of the circumscribing circles for various values of n, the number of sides of the regular n-gon. I think it is clear that one cannot say with certainty, looking at the numbers, whether the sequence converges or not.

TABLE 1 RADII OF CIRCUMSCRIBING CIRCLES

n	R_n
1000	8.6572
2000	8.6786
3000	8.6857
4000	8.6893
5000	8.6915
6000	8.6929
7000	8.6939
8000	8.6947
9000	8.6952
10000	8.6957

For convenience define R_2 to be 1. Elementary geometry gives

$$R_n = R_{n-1}/\cos(\pi/n), \qquad n > 2,$$

or

$$\ln R_n - \ln R_{n-1} = -\ln\cos(\pi/n),$$

and summing from 3 gives

$$\ln R_n = -\sum_{k=3}^{n} \ln \cos(\pi/k).$$

Now examination of the first 3 terms of the Maclaurin series for $-\ln\cos(x)$ shows

$$-\ln \cos(\pi/k) = \frac{\pi^2}{2k^2} + o(1/k^3) \qquad \text{as } k \to \infty.$$

Since the sum of $1/k^2$ converges, we see that the sequence of R_n does in fact converge. But now the question becomes, "To what does it converge?" That question is more difficult to answer, and modern technology can be very helpful. We might note, as an aside, that neglecting the $o(1/k^3)$ term in $-\ln\cos(\pi/k)$ and using $\Sigma_{k\geq 1} 1/k^2 = \pi^2/6$, produces $\lim_{n\to\infty} R_n \geq 7.02$, which is not very helpful; more on this later.

If $y = -\ln\cos(x)$, $y'' = \sec^2(x)$; hence $\ln\cos(1/t)$ is convex (lies above any tangent). Replacing $f(x)$ in the integral below by $f(k) + f'(k)(x-k) + f''(\phi)/2(x-k)^2$ and integrating gives

$$\int_{k-1/2}^{k+1/2} f(x)\,dx = f(k) + \tfrac{1}{24}f''(\theta) \qquad k - \tfrac{1}{2} < \theta < k + \tfrac{1}{2},$$

after an application of the mean value theorem for integrals. This is, of course, the midpoint formula of numerical quadrature with error term, so when $f'' \geq 0$ the integral exceeds the function evaluation. Consequently,

$$\sum_{k\geq n} f(k) \leq \int_{n-1/2}^{\infty} f(x)\,dx,$$

or in this case

$$\sum_{k\geq n} -\ln\cos(\pi/k) \leq \int_{n-1/2}^{\infty} -\ln\cos(\pi/x)\,dx.$$

This last integral can be evaluated using a good mathematics package; some require reformulation to work effectively, viz. putting $t = \pi/x$. For $n = 10001$ the integral is then found to have the value .000493455660681127.

Putting this together with R_{10000}, we have

$$R_\infty < R_{10000} * \exp(.000493455660681127) = 8.70003662617486.$$

To obtain a lower bound, observe that

$$\int_{k}^{k+1} f(x)\,dx = \tfrac{1}{2}f(k) + \tfrac{1}{2}f(k+1) - \tfrac{1}{12}f''(\phi) \qquad k < \phi < k+1.$$

This is the classical trapezoidal rule with error term.[1] If $f'' > 0$, the integral is smaller than the sum of the first two terms on the right, so

$$\sum_{k\geq n} f(k) \geq \int_{n}^{\infty} f(x)\,dx + \tfrac{1}{2}f(n).$$

[1]Suppose $f(1) = f_1 = 0$, and $f(0) = f_0$. Then an integration by parts followed by a replacement of $f'(x) = f'(\frac{1}{2}) + (x - \frac{1}{2})f''(\phi)$ yields $\int_0^1 f dx - \frac{1}{2}(f_1 + f_0) = -\int_0^1 (x - \frac{1}{2})f'\,dx = -f''(\theta)/12$, $0 < \theta < 1$, after an application of the mean value theorem for integrals. Finally, replacing an arbitrary $f(x)$ with $f(x) - f_1$ produces the desired result.

With the choice $f(x) = -\ln\cos(\pi/x)$ the two terms on the right are

$$.000493430947725898 + \tfrac{1}{2} * 4.93381546687769 * 10^{-8},$$

and we have

$$R_\infty > R_{10000} * \exp(.000493455616803233) = 8.70003662579312.$$

So $R_\infty = 8.700036626$ accurate to 9 places.

A referee points out that if you have Mathematica at your disposal, you can obtain a given number of terms in the power series around $x = 0$ for $-\ln\cos(\pi/k)$ and then sum from 3 to ∞ on k. This makes use of the fact that Mathematica can produce the algebraical expressions for power series, the system knows the values of $\zeta(2k)$ for arbitrarily high k, and it can work to arbitrarily high precision. You must of course still produce upper and lower bounds to ensure the accuracy of your proposed answer.

Now the purpose of this example was not to obtain the specific answer to a specific problem but to illustrate a method of some general utility. This latter method certainly will work for this particular problem and work well, but it seems to me not to possess the desired general applicability.

Decimation of the Harmonic Series. As a second example, we selectively mutilate the harmonic series by removing all those terms whose denominators have a zero in their decimal representation. Thus

$$s = 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{9} + \tfrac{1}{11} + \cdots + \tfrac{1}{99} + \tfrac{1}{111} + \cdots$$

First, does the series converge? There are 9 terms with a single digit denominator, each at most 1, there are 9^2 terms having a 2 digit denominator, each less than $\frac{1}{10}, \ldots$. Hence

$$s < 9 * 1 + 9^2 * \tfrac{1}{10} + 9^3 * \left(\tfrac{1}{10}\right)^2 + \cdots = 90,$$

but how much less than 90 is s?

Series convergence is painfully slow. R. P. Boas points out [1] that a sum of 10^{18} terms leaves a remainder greater than 1, and even that supposes you could do the arithmetic without roundoff. It seems that brute force computing will not supply the answer in any easy manner, and we are in need of a reasoned attack on the problem.

A refinement of the preceding argument is the following. If N^- denotes the natural numbers without a zero in their decimal development and $A = N^- \cap \{k\colon 11111 \le k \le 99999\}$, then

$$\eta = \sum_{k \in A} 1/k = 1.474621200429149\ldots .$$

Because of the way floating point arithmetic is done on most machines, it is best to add small numbers to small numbers, never small to large. Consequently η is evaluated by summing backwards from 99999 to 11111. This is yet another example of thoughtful human intervention.

If $\zeta = \sum_{k \in B} 1/k$, where $B = N^- \cap \{k\colon 1 \le k < 11111\}$, an upper bound, ub, may be found as before:

$$ub = \zeta + \eta + \eta * \tfrac{9}{10} + \eta\left(\tfrac{9}{10}\right)^2 + \cdots = 23.103719216.$$

To get a lower bound, lb, one modifies the above argument slightly. Integers in N^- of the form $11111x$ are at most 111119, those of the form $22222x$ are at most $222229, \dots,$ those of the form $99999x$ are at most 999999. But

$$\frac{1}{111119} = \frac{1}{111110} * \frac{111110}{111119} < \frac{1}{111110} * .99991$$

$$\vdots$$

$$\frac{1}{999999} = \frac{1}{999990} * \frac{999990}{999999} > \frac{1}{999990} * .99991.$$

The sum of these terms is therefore bounded below by $\eta * (\frac{9}{10}) * .99991$. The next decade is bounded below by $\eta * (\frac{9}{10})^2 * .99991$, etc. Finally, then

$$lb = \zeta + \eta + \left\{ \eta * \frac{9}{10} + \eta * \left(\frac{9}{10} \right)^2 + \cdots \right\} * .99991 = 23.1025,$$

and $23.1025 \le s \le 23.1037$.

Riemann's theorem on conditional convergence. A nice application of modern technology may be made to the summing of conditionally convergent series.

$$1 - \frac{1}{2} + \frac{1}{3} - \cdots - \frac{1}{2n} = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{n+n}$$

$$= \left(\frac{1}{1 + \frac{1}{n}} + \cdots + \frac{1}{1 + \frac{n}{n}} \right) \cdot \left(\frac{1}{n} \right)$$

The limit on the left as $n \to \infty$ is

$$\sum_{k \ge 1} (-1)^{k+1}/k,$$

while on the right it is

$$\int_0^1 \frac{1}{(1+s)} ds = \ln 2.$$

Now the sum of the positive terms diverges to $+\infty$, while the sum of the negative terms gives $-\infty$. To get a rearranged series to converge to, say, 2 we take enough positive terms in order of decreasing size to just exceed 2, then enough negative terms in order of size to just fall below 2, then repeat the process indefinitely. Figure 2 shows a plot of the resulting partial sums of this rearranged series. Such a figure is very unlikely to appear without modern computing equipment, for the amount of arithmetic required is just too painful to be worthy of the effort.

The logarithmic increase in the sum of positive terms is immediately obvious from the plot, as is the gradual diminution of the individual terms. It is also evident how few negative terms are necessary to bring the sum below 2. The plot is illuminating and something is suggested: it would seem that a rearrangement of this familiar series, which converges to something other than $\ln 2$, must have terms

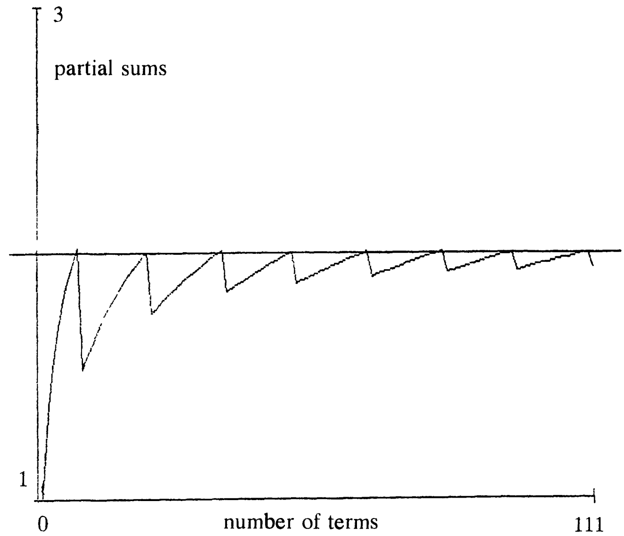


Figure 2. Rearrangement converging to 2.

arbitrarily far removed from their original position. In fact, one can prove the following

Theorem. *If*

(1) $$\sum_{i\geq 1} a_i = s \qquad \textit{conditionally convergent}$$

(2) $$\sum_{i\geq 1} a_{f(i)} = t \neq s \quad \textit{a convergent rearrangement of } (1),$$

then

$$\textit{Given } N\ \exists k \ni : |k - f(k)| > N.$$

Proof: Suppose not. Then there exists an N such that $|k - f(k)| \leq N$ for all k. Let

$$\sum_{i=1}^{n} a_{f(i)} = \sum_{i=1}^{n+N} a_i - R_n,$$

where $R_n = \sum_{j'=1}^{N} a_{j'}$. But for each n there exists an $M(n)$ such that $k > M(n) \Rightarrow f(k) > n$, and hence the N values j' increase with n. Therefore $R_n \to 0$, but the left-hand side tends to t, and the right-hand side tends to $s + 0$, which is impossible.

It seems unlikely that this fact is unknown, but I have not been able to locate a source. The result may not be earth-shaking, but its origin illustrates an important point. Modern technology often suggests possibilities that might well have otherwise escaped attention. It is not solely a source of verification; present-day computing allows prospecting for results not only in data analysis but in mathematics proper.

Rate of convergence. A concluding example asks the following. Which of the following two series converges faster

$$\sum_{n\geq 1} 1/n^2 \quad \text{or} \quad \sum_{n\geq 1} (-1)^{n+1}/n^2?$$

The first converges, as Euler showed, to $\pi^2/6 = 1.644934\ldots$, the second to $\pi^2/12$. Now

$$\sum_{n \geq N} 1/n^2 < \int_{N-\frac{1}{2}}^{\infty} 1/x^2 \, dx = \frac{1}{N - \dfrac{1}{2}},$$

and using the usual estimate for the error in an alternating series (the absolute value of the first term neglected)

$$\left| \sum_{n>N} (-1)^{n+1}/n^2 \right| \leq \frac{1}{(N+1)^2}.$$

Thus it appears the alternating series is the more accurate for a given number of terms.

But examine

$$e^{-x} = 1 + x + x^2/2! + x^3/3! + \cdots,$$

and

$$e^{-x} = 1 - x + x^2/2! - x^3/3! + \cdots.$$

If we sum the first N terms of each, the relative error of the first is

$$1 - \exp(-x) \sum_{k=0}^{N-1} x^k/k!,$$

while that of the second is about

$$\frac{x^N}{N! \exp(-x)}.$$

For even moderate x, the second is worse than the first: $x = 2$, $N = 5$, yields .0527 for the first and 1.97 for the second. The conclusion is, I believe, that you must think about the purpose to which you wish to put an evaluated process. Most often it is relative error that is the thing of importance. Blind arithmetic will surely lead one astray. On the other hand, a carefully-chosen numerical example may suggest an avenue of attack heretofore unenvisioned.

The best dictum of which I am aware is that of R. W. Hamming [**2**], "The purpose of computing is insight, not numbers!"

REFERENCES

1. R. P. Boas, Jr., Partial Sums of Infinite Series and How They Grow, *Amer. Math. Monthly* 84 (1977) 237–258.
2. R. W. Hamming, *Numerical Methods for Scientists and Engineers*, 2nd edition, Dover, 1986.
3. J. R. Newman, *The World of Mathematics*, Simon and Schuster, New York, 1956.

Department of Mathematical Sciences
Stevens Institute of Technology
Castle Point on Hudson
Hoboken, NJ 07030
rpinkham@stevens-tech.edu

Tartaglia's Inverse Problem in a Resistive Medium

C. W. Groetsch

Mechanics is the vehicle of all physical theory. Mechanics is the vehicle of war.
The two have been inseparable.

David Oliver, *The Shaggy Steed of Physics*

1 INTRODUCTION. In late renaissance Italy, the practical demands of warfare and the emerging science of mechanics converged to give birth to one of the earliest mathematicophysical theories: the ballistics of projectiles in a nonresistive medium. "The Newly Discovered Invention of Nicolò Tartaglia of Brescia, Most Useful for Every Theoretical Mathematician, Bombadier, and Others, Entitled *New Science*" (Venice, 1537) was a groundbreaking, though flawed, treatment of primitive ballistics that nearly did not see the light of day. In his dedication of the work Tartaglia avers that "I fell to thinking it a blame-worthy thing, ... deserving no small punishment by God—to study and improve such a damnable exercise, destroyer of the human species.... For which reason... not only did I wholly put off the study of such matters and turn to other studies, but also I destroyed and burned all my calculations and writings that bore on this subject.... But now, seeing that the wolf is anxious to ravage our flock, while our shepherds hasten to the defense, it no longer appears permissible to me at present to keep these things hidden." [**2**]. Tartaglia was no stranger to the horrors of war. The nickname by which we know him ("the stutterer") was the result of a sword wound to the throat he suffered at age twelve during a sacking of Brescia.

The "wolf" to which Tartaglia refers was Suleiman the Magnificent (1496?–1566) (as the *tenth* Ottoman Sultan, Suleiman was also known as the "Perfector of the Perfect Number" [**4**]). Suleiman had taken Belgrade, burned Buda, laid siege to Vienna, and was a clear and present threat to Venice and her commercial interests. The parallel between Tartaglia's moral struggle and that faced by some scientists on both sides of the Atlantic at the dawn of the nuclear age is inescapable (e.g., [**5**], [**7**]).

In 1531, Tartaglia devised the gunner's square (essentially a carpenter's square with attached protractor dividing the right angle into 12 "punti", or points, which were read off with a plumb line; see Figure 1) making fairly accurate determination of the angle of elevation of a field piece possible. Although Tartaglia in his theoretical studies did not get the curve of the projectile motion right, he did declare that the maximum range is attained at an elevation of 45° (6 punti) [**2**]. Moreover, in the letter of dedication of his *New Science* to Francesco Maria della Rovere, the Duke of Urbino, he remarks, "I knew that a cannon could strike in the

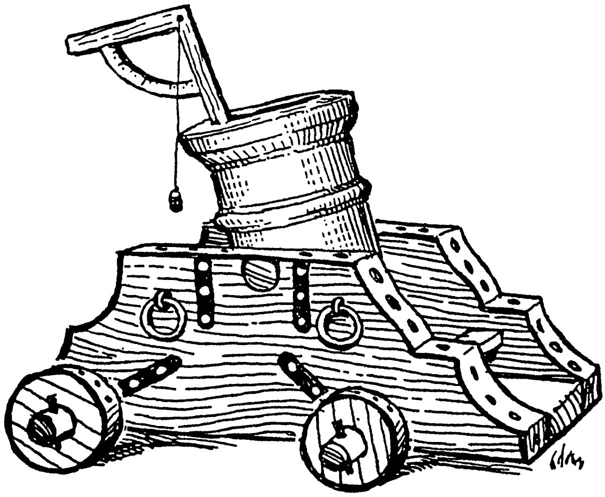


Figure 1. Tartaglia's Square

same place with two different elevations or aimings, I found the way of bringing this about, a thing not heard of and not thought by any other, ancient or modern." [**2**].

A century later, Galileo did get the curve right. His great insight was that the motion of the projectile can be decomposed into horizontal and vertical components, the horizontal motion having constant velocity and the vertical motion being constantly accelerated by gravity. Having previously established that distance is proportional to the square of the time in constantly accelerated motion, he then deduced (by strictly Euclidean arguments) that the trajectory of a projectile in a nonresistive medium is parabolic. Moreover, he gave a proof that the range is a maximum when the angle of elevation is "at the sixth point of the [gunner's] square," i.e., at 45°. Galileo goes on, in his customary style of dialogue, to discuss (without attribution to the Brescian mathematician) the surprising fact that had already been noticed by Tartaglia, namely that the inverse problem of determining elevation angles leading to a given submaximal range has precisely two solutions. He puts his proposition into the mouth of one of the interlocutors, Salviati: "having gained by demonstrative reasoning the certainty that the maximum of all ranges of shots is that of elevation at half a right angle, the Author demonstrates to us something that has perhaps not been observed through experiment; and this is that of all other shots, those are equal [in range] to one another whose elevations exceed or fall short of half a right angle by equal angles." [**3**].

To show that the range, for a given muzzle velocity v and angle of elevation θ, is given by

$$R = \frac{v^2}{g}\sin 2\theta,$$

where g is the gravitational acceleration constant, is a fairly common exercise in calculus texts. From this, it follows that the maximal range occurs at $\theta = \pi/4$. It is much less common to find in texts the inverse problem of determining the angles of elevation for a given range.

This note concerns Tartaglia's inverse range problem for projectiles in a medium that resists motion in proportion to the projectile's velocity. It turns out that the simple *linear* differential equations that model resistance proportional to the velocity lead to an essentially *implicit* expression for the range. In this case, an indirect approach, using some elementary analysis related to a fixed point equation, shows that each submaximal range is attained for exactly two elevations. Our treatment is meant to provide an illustration of some interesting ideas from elementary analysis and addresses only the question of the existence of exactly two solutions of the inverse problem; the much more difficult question of the exact relationship between the solutions is not considered.

2 THE LINEAR MODEL. Suppose the projectile has unit mass and the medium resists motion in proportion to the velocity, with constant of proportionality $k > 0$. The equations of motion are then

$$\ddot{x} = -k\dot{x}$$
$$\ddot{y} = -g - k\dot{y}$$

(With apologies to Babbage, we follow the physical community in using Newtonian dots for time derivatives; a derivative with respect to a nontemporal variable will be indicated by the conventional prime.) If the muzzle velocity is v, the angle of elevation is θ, and the projectile is initially at the origin, then the equations of motion have the unique solution

$$x(t) = (v\cos\theta)(1 - e^{-kt})/k$$
$$y(t) = \left(\frac{v\sin\theta}{k} + \frac{g}{k^2}\right)(1 - e^{-kt}) - \frac{gt}{k}.$$

Eliminating t from the first equation and setting $y = 0$ (for the impact point) we find, after a little rearranging, that the range $R = R(\theta)$, for a given elevation θ, satisfies

$$R = \frac{\cos\theta}{a}(1 - e^{-A(\theta)R}), \tag{1}$$

where $A(\theta) = a\sec\theta + b\tan\theta, a = k/v, b = k^2/g$. For fixed $k > 0, v > 0, \theta \in [0, \pi/2]$, equation (1) shows that the range is a fixed point of the function

$$f(s) = \frac{\cos\theta}{a}(1 - e^{-A(\theta)s}). \tag{2}$$

This bird is more easily caged if we trim some of its plumage by writing it in the form

$$f(s) = c(1 - e^{-ds}) \qquad (s \geq 0) \tag{3}$$

where $c = (\cos\theta)/a$ and $d = A(\theta)$. We note that for $\theta \in (0, \pi/2)$, $c > 0$ and $cd = 1 + b/a\sin\theta > 1$. Some simple properties of f that we will need, most of which are evident from a quick sketch, are gathered in the following proposition.

Proposition 1. *If $c > 0$ and $cd > 1$, then the function f in (3) has a unique positive fixed point p. Moreover,*

(i) $0 < s < p$ *if and only if* $f(s) > s$ *and* $p < s$ *if and only if* $f(s) < s$, *and*
(ii) $0 < f'(p) < 1$.

Proof: It is clear that $f\colon [0, c] \to [0, c)$, $f(0) = 0$, and $f'(0) = cd > 1$. Hence $f(s) > s$ for all sufficiently small positive s and $f(c) < c$. Therefore, f has a fixed point in $(0, c)$. Were there another positive fixed point, then, since 0 is a fixed point, f would have three fixed points and f'' would have a zero. However, $f''(s) < 0$ for all s.

The statements in (i) follow immediately from the fact that $f'(0) > 1$, $f(c) < c$, and $(0, p) \cup (p, \infty)$ is free of fixed points.

From (3) and the fact that p is a fixed point, we observe that

$$f'(p) = cde^{-dp} = cd - pd > 0. \tag{4}$$

Using the fact that $e^x > 1 + x$ for $x > 0$, we get

$$(x + 1)(1 - e^{-x}) > x.$$

Setting $x = cd - 1 > 0$, we find that

$$f\left(\frac{cd - 1}{d}\right) = c(1 - e^{1-cd}) > \frac{cd - 1}{d}.$$

Therefore, $p > (cd - 1)/d$, by (i). From (4) we then conclude that $f'(p) < 1$. □

The proposition ensures that fixed point iteration converges monotonically to the positive fixed point of (3) for any initial approximation $s_0 > 0$. In particular, we see that the projectile range $R(\theta)$ may be computed by fixed point iteration for any $\theta \in (0, \pi/2)$. For example, the computed range function $R(\theta)$ (with $v = 100$ and $g = 32.2$) is plotted (courtesy of MATLAB [6]) in Figure 2 for several values of the resistance constant k. It is evident from the plots that Tartaglia's observation (to each submaximal range value there correspond two distinct elevations) continues to hold in the resistive medium. A proof of this is given in the next section.

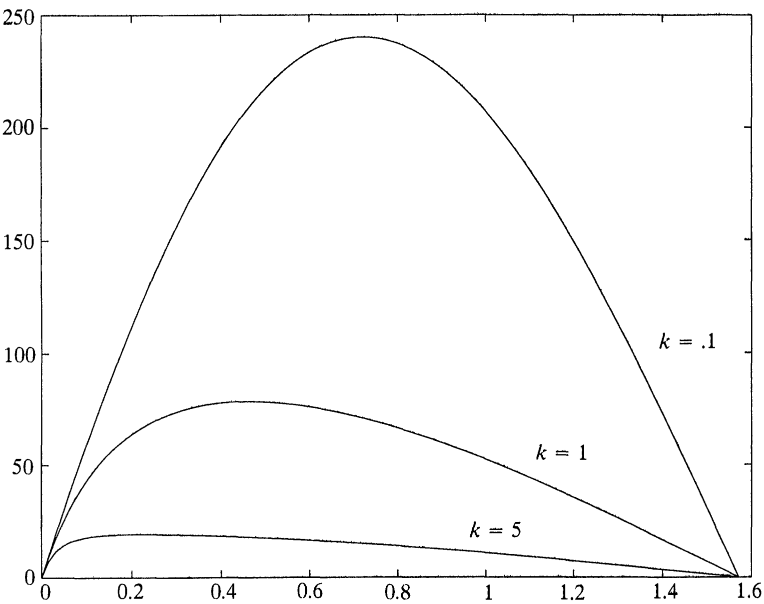


Figure 2. Range vs. Angle of Elevation

3 THE INVERSE PROBLEM. As noted in the previous section, the range function satisfies

$$\begin{aligned} R(\theta) &= \frac{\cos\theta}{a}\left(1 - e^{-A(\theta)R(\theta)}\right) \\ &= F(\theta, R(\theta)) \end{aligned} \tag{5}$$

where

$$F(\theta, r) = \frac{\cos\theta}{a}\left(1 - e^{-A(\theta)r}\right).$$

Since $(\cos\theta/a)A(\theta) = 1 + (b/a)\sin\theta > 1$ for $\theta \in (0, \pi/2)$, the proposition ensures that a unique $R(\theta)$ exists. We now show that the function R is differentiable on $(0, \pi/2)$ (in fact, existence, uniqueness, and differentiability of R may be obtained via the implicit function theorem **[1]**, but as we have already established existence and uniqueness, we give a direct argument for differentiability). For $\theta \in (0, \pi/2)$ and $|\Delta\theta|$ sufficiently small, we have from (5)

$$\begin{aligned} \Delta R &= R(\theta + \Delta\theta) - R(\theta) \\ &= \frac{\partial F}{\partial \theta}(\theta + q\Delta\theta, R + q\Delta R)\Delta\theta + \frac{\partial F}{\partial r}(\theta + q\Delta\theta, R + q\Delta R)\Delta R \end{aligned} \tag{6}$$

for some $q \in (0, 1)$. However, from (ii) of the proposition

$$\frac{\partial F}{\partial r}(\theta, R(\theta)) = f'(R(\theta)) < 1,$$

where f is given by (3). Therefore, since F has continuous partial derivatives, it follows from (6) that R is differentiable and

$$R'(\theta) = \frac{\partial F}{\partial \theta}(\theta, R(\theta)) \Bigg/ \left(1 - \frac{\partial F}{\partial r}(\theta, R(\theta))\right). \tag{7}$$

Lemma 1. *Suppose* $A(\theta) = a\sec\theta + b\tan\theta$, $B(\theta) = b\sec\theta + a\tan\theta$ *and* $R(\theta)$ *is defined by* (1). *Then*

(i) $A'(\theta) = B(\theta)\sec\theta$, $B'(\theta) = A(\theta)\sec\theta$,

(ii) $A(\theta)^2 - B(\theta)^2 = a^2 - b^2$, *and*

(iii) $1 - e^{-A(\bar\theta)b/B(\bar\theta)a} = b\sec\bar\theta/B(\bar\theta)$, *if* $R'(\bar\theta) = 0$.

Proof: The first two parts are straightforward. If R is defined by (1), and $R'(\bar\theta) = 0$, then by (7)

$$0 = -\frac{\sin\bar\theta}{a}\left(1 - e^{-A(\bar\theta)R(\bar\theta)}\right) + \frac{\cos\bar\theta}{a}e^{-A(\bar\theta)R(\bar\theta)}A'(\bar\theta)R(\bar\theta),$$

which, by use of (1) and (i), simplifies to

$$a\tan\bar\theta = B(\bar\theta)e^{-A(\bar\theta)R(\bar\theta)}.$$

Solving (1) for $e^{-A(\bar\theta)R(\bar\theta)}$ and substituting, this gives

$$\frac{B(\bar\theta) - a\tan\bar\theta}{aB(\bar\theta)} = R(\bar\theta)\sec\bar\theta$$

or

$$R(\bar{\theta}) = \frac{b}{aB(\bar{\theta})}.$$

Substituting this into (1) gives (iii). □

We now have sufficient weapons for an indirect attack on the inverse problem.

Proposition 2. *The equation $R(\theta) = r$ has precisely two solutions $\theta \in [0, \pi/2]$ for each nonnegative number r smaller than the maximum range.*

Proof: Since R is a nonnegative continuous function with $R(0) = R(\pi/2) = 0$, there are at least two elevations θ that result in a given submaximal range. If there were more than two such elevations, then there would be two values, $\theta_1 \neq \theta_2$ with

$$R'(\theta_1) = R'(\theta_2) = 0.$$

Then equation (iii) of the Lemma would have two distinct solutions in $(0, \pi/2)$ and hence for some $\theta \in (0, \pi/2)$ we would have

$$\left(1 - e^{-A(\theta)b/B(\theta)a}\right)' = \left(\frac{b \sec \theta}{B(\theta)}\right)'$$

or, equivalently

$$e^{-A(\theta)b/B(\theta)a}\left(A'(\theta)B(\theta) - A(\theta)B'(\theta)\right) = a \sec \theta\left(B(\theta)\tan \theta - B'(\theta)\right).$$

Now, using (i) and (ii) of the Lemma, we find that for some $\theta \in (0, \pi/2)$

$$e^{-A(\theta)b/B(\theta)a}\left(1 - (b/a)^2\right) = 1,$$

which is impossible. □

ACKNOWLEDGMENTS. This work was supported by the National Science Foundation and the Charles Phelps Taft Memorial Fund.

REFERENCES

1. R. Courant and F. John, *Introduction to Calculus and Analysis*, Vol. II, Springer-Verlag, New York, 1989.
2. S. Drake and I. E. Drabkin, *Mechanics in Sixteenth-Century Italy*, University of Wisconsin Press, Madison, 1969.
3. G. Galilei, *Two New Sciences* (Elzevirs, Leyden, 1638), translated, with a new introduction and notes, by Stillman Drake, Second Edition, Wall and Thompson, Toronto, 1989.
4. J. P. D. B. Kinross (Lord Kinross), *The Ottoman Centuries*, Morrow, New York, 1977.
5. W. Lanouette, *Genius in the Shadows: A Biography of Leo Szilard, the Man Behind the Bomb*, University of Chicago Press, Chicago, 1992.
6. MATLAB, 1989, The Mathworks, Inc., 21 Elliot St., South Natick, MA 01760.
7. T. Powers, *Heisenberg's War: The Secret History of the German Bomb*, Knopf, New York, 1993.

Department of Mathematics
University of Cincinnati
Cincinnati, OH 45221-0025
groetsch@uc.edu

The Current State of Actuarial Science

Donald P. Minassian

Little in the literature describes actuarial science to general mathematicians. No full *Monthly* article appears on this topic. Two short papers since 1980 are a $2\frac{1}{2}$ page teaching note [**3**] and a $1\frac{1}{2}$ page book review [**2**]. Much has changed since then. Our goal is to present non-technical, current material on actuarial organizations, employment, examinations and education, research, and "practical" issues.

1. ACTUARIAL ORGANIZATIONS. There are several actuarial organizations in North America. Most deal with education, examinations, research, and advancing the profession. The largest in the world is the Society of Actuaries (SOA) with some 16,000 members, handling life and health insurance and pensions. About 20% of its members are from Canada, 8% from outside North America, and the rest from the U.S. The Casualty Actuarial Society (CAS), with membership about 2,300, handles casualty insurance like auto and homeowners. The Canadian Institute of Actuaries (CIA—the initials do confuse!), with some 1,900 members, handles general insurance for Canadian actuaries, with much overlap in members, issues and examinations with the SOA.

To belong to either the SOA or CAS requires passing many difficult examinations leading at least to Associateship (details later). The SOA/CIA examinations duplicate those of the CAS in the earliest stages. Membership in the American Academy of Actuaries (AAA), now some 12,000 strong, requires Associateship in another society and three years practical work. The CIA has like requirements, but the experience must span $1\frac{1}{2}$ years of Canadian work in the last three years. Also, Canadian versions of some examinations must be taken. Full membership in the SOA, CAS, or CIA requires Fellowship (details later). Smaller groups are the American Society of Pension Actuaries and the Conference of Consulting Actuaries, plus foreign actuarial societies and a world body. Also, some cities in North America have clubs. Disciplinary muscle is provided by the AAA and the CIA, which enact standards of professional conduct and punish violators. In 1988 an Actuarial Standards Board arose to prepare and initiate public hearings on actuarial standards, enacting new ones and reviewing those already existing.

All told, there are at most 20,000 U.S. actuaries with at least Associateship. Contrast this number with some 325,000 in the American Institute of Certified Public Accountants and about 865,000 U.S. lawyers!

I'd like to make some additional comments about the SOA since it is largest and I know it best. The count of new FSAs (Fellows, all exams passed) has risen markedly. From 1889 to 1948 there were 897 new FSAs, but from 1949 to 1994 there were 7,690. Of the 7,100 FSAs in January 1993, 871 became so in the 1990s, two of seven after 1984, about 50% after 1979, 70+% after 1974, and 80% since 1969.

The SOA has a President, who serves a year, and a Board of Governors. Many SOA Committees handle research, education and examinations, membership ser-

vices, etc. Most committee members, as well as the President and Board, are part-time volunteers who maintain regular jobs in industry. But the SOA has 18 full-time officers including 8 FSAs and 2 Ph.Ds, and some 50 support staff, at offices near Chicago.

2. EMPLOYMENT. Actuaries use mathematics and other knowledge (see Section 3) to handle practical problems in insurance and related areas.

Of the 16,000 SOA members, about 7,200 work for insurance companies, 5,500 are consultants who work for themselves or consulting firms as advisors to industry and government in work such as helping to shape employee benefit plans, 400 are in state or provincial insurance departments, 180 are in colleges or universities, 1,100 are retired, 800 are unaffiliated, with the rest "miscellaneous" (11/1/94). The last two categories are growing as actuaries augment their traditional roles, as described later.

In the two largest fields—insurance companies and consultants—work lies in many areas. Examples are:

Pricing: what premiums to charge;
Product development: creating new types of insurance policies;
Reserving/valuation: fixing what an insurer must retain to pay future claims;
Studying rates of mortality or morbidity or other loss and then projecting future claims;
Reinsurance: one company farms out risky or excess business to another, often larger company to dampen unfavorable variance in claims;
Dividends: how much premium is safe to return to policyholders;
Mergers and acquisitions: what one company should pay to acquire another;
Working with the investment department;
Auditing: testing often with CPAs the reasonableness of financial statements;
Modelling: predicting the future under various scenarios;
Pension work;
Top management (thus abandoning many strictly "actuarial" duties).
Of mounting importance is general finance, exceeding traditional roles.

It is safe to say that most actuaries majored in mathematics, actuarial science, or related quantitative areas in college. The mathematics used in most practical actuarial work is elementary algebra, calculus, basic probability and statistics, numerical analysis, operations research, and "actuarial" subjects such as interest theory, life contingencies, or contingencies of ill health, accident or fire for health or casualty work. Based on my life and casualty work, most of the mathematics is elementary, but deeper knowledge can help. The mathematics in more advanced research is discussed in Section 4. Finally, as everyday work grows more exacting, more and more academic research will find its way into the workplace, e.g., finance theory.

3. EXAMINATIONS AND EDUCATION. To become a full actuary requires passing many difficult examinations. In both the SOA and the CAS one first becomes an Associate and then a Fellow.

A key difference between actuarial exams and many other professional exams is that most of the former are taken *on the job*; candidates already have full-time employment. Another difference is that the process usually takes eight to ten years, while in law there is one qualifying exam (the bar exam) after three years of

law school, and in accounting there is one $2\frac{1}{2}$ day CPA exam after college, although the exam may be passed in parts.

Because a prospective actuary is employed full-time, dedication to study must be severe. Although some study time is usually given by employers, most study must be done in one's "leisure." Exams are given in May and November (the first two also in February) with eight months of the year usually devoted to study. College courses help with some lower exams (see the following list of exams).

Does work on the job help to pass exams? Only marginally, for at least three reasons: (1) many exam questions probe for knowledge deeper than that used on the job, (2) a practicing actuary usually works in just a few areas, but one is tested in many areas, (3) sometimes the "practical" solutions on the job are not the answers examiners seek.

The examinations are not research-oriented. The purpose is to produce "practical" people. Few PhDs, of which there are but a handful in actuarial science as compared to mathematics, are involved in composing exam questions, which are written by committees of practicing actuaries.

The actuarial exams I took had a marked memorization component, especially on upper exams, although the SOA may lately be composing more creative questions that require one to apply knowledge to "original" situations. Emphasis on memorization could discourage mathematicians. All the exams are closed book.

Also, in contrast to at least some other professions, specific study materials (rather than simply the topics to be learned) are generally prescribed. Both the SOA and the CAS prepare study notes upon which exam questions are based. Of course, such prescription of study materials is consonant with need for memorization.

Time pressure on exams is strong. Although I have taken exams for the CPA and the law, as well as doctoral mathematics exams, I have rarely been so rushed as on my SOA exams, particularly the lower, mathematical ones, many of which I failed to complete.

Another difference between actuarial and some other examinations is the low pass ratios. Roughly 40% of candidates pass any given exam, although exams may be repeated, depending on the patience of the candidate (and perhaps of the employer).

A final contrast to at least some professions is the huge reliance on unpaid volunteers for preparing, administering, and grading exams and for writing study notes. These volunteers often take time away from their full-time, paid jobs. I think quality may have suffered as a result. For example, I feel SOA study materials could be clearer, more correct, less duplicative and/or contradictory, etc. Most authors lack teaching experience, have little chance to interact with other authors, may not even know what these other authors write, and must meet time deadlines while trying to do decent work at their regular jobs. The SOA and CAS owe an enormous debt to these selfless volunteers, and it is a wonder things are as good as they are, but quality can be improved.

I now deal in more detail with the SOA exams. One first becomes Associate of the Society of Actuaries (ASA), then Fellow (FSA).

Things have changed since [**2**] noted the "ten-part Fellowship exams." Those parts are now split into smaller pieces, called "Courses." Now one needs 20 to 30 exams, or "Courses," for the FSA, based on choice of electives, since Courses may carry differing credits. The various *tracks*, described below, can also make a difference. All this is called FES for Flexible Education System. The CAS still keeps its old 10-part exams.

Until July, 1995, 200 credits sufficed for the ASA and 250 more for the FSA. Usually 10 credits equal an exam-hour, so 450 total credits yield some 45 exam-hours. Generally, ASA exams are mathematical, or at least "quantitative," while upper exams are much less so.

Save for two 1-hour elective "intensive seminars", where students meet for a week under SOA tutelage, the ASA exams are machine-graded, multiple-choice. Later exams are largely "written answer"—mostly short essays. Some 55,000 individual exams are given annually. Once one completes 300 credits, there is a choice of four tracks: *individual* insurance/annuities, *group* insurance/annuities, *pensions*, or a new *finance* track. A track has some 90 credits and, finally, one picks 60 credits of electives from any track. Save for the two intensive seminars and the Fellowship Admissions Course, the SOA gives no "in-house" exam training, although it issues much study material.

In July, 1995, 100 credits were moved to the ASA level, although the total of 450 credits for the FSA remains. The moved credits were the so-called Core FSA Courses, which one takes before specializing in a track. Here is a list of the ASA exams:

	Course No.	Description	Credits
A) Required:			
	100	Calculus and Linear Algebra	30
	110	Probability and Statistics	30
	120	Applied Statistical Methods	15
	140	Mathematics of Compound Interest	10
	150	Actuarial Mathematics	40
	151	Risk Theory	15
	160	Survival Models and Construction of Tables	15
		Total credits	155
B) Electives:	(choose 45 credits)		
	130	Operation Research	15
	135	Numerical Methods	10
	CAS Part 4B	Credibility Theory and Loss Distributions	20
	121	Intensive Seminar on Applied Statistical Methods	10
	152	Intensive Seminar on Risk Theory	10
	161	Mathematics of Demography	10
	165	Mathematics of Graduation	10
C) Additionally required for ASA after July, 1995:			
	200	Introduction to Financial Security Programs	30
	210	Introduction to Actuarial Practice	25
	220	Introduction to Asset Management and Corporate Finance	30
	230	Principles of Asset/Liability Management	15
		Total credits	100

In the 100-Courses, "Actuarial Mathematics" deals with contingencies of survival and other decrements, while "Graduation" smooths mortality and other decrements using mathematical formulas and techniques. A "quasi-track" exists at ASA level: to become an "Enrolled Actuary", licensed by the U.S. to do pension work, requires only two exams. The first, quantitative one, called EA-1, tests mathematics of compound interest plus actuarial mathematics, while the nonquantitative EA-2 tests topics such as pension law. So here is a quick way to become titled!

Of nearly 300 examination centers, some 60 are abroad, even in China, with the rest in North America. Examinations are in English, except for Canada, where they are also in French (even outside Canada some post-ASA exams are bilingual).

Recently, the SOA has given 30 elective credits for writing a research paper on actuarial science; this is roughly a master's thesis and is the single research-oriented area in the examination system. Through late 1994 over forty proposals for research papers have been submitted, of which about half have been approved with eight papers actually accepted. With 55,000 exams given yearly, clearly very few prospective actuaries use this option.

Once all examinations are passed, one step remains—also rather new. Candidates attend a $2\frac{1}{2}$-day seminar called a "Fellowship Admissions Course" or FAC. Candidates deal with (1) ethical issues and (2) case studies, which are practical situations, sometimes from real life, needing solution. Almost all FAC attendees pass, then receive their FSA diploma at a banquet.

Course 100 (Calculus and Linear Algebra) changed in November, 1995. Keeping the overall time limit, the number of questions dropped from 60 to 45, and a short table of integrals was supplied. The aim is to stress concepts over computation skills. Also, the text **[1]** for the meatiest ASA exam, Course 150, is being revised to (1) relax the assumption that people die independently, a relaxation that complicates the statistical treatment, (2) introduce stochastic rather than fixed interest rates, which is also complicating, and (3) widen "contingencies" to other business situations such as lives of light bulbs, default rates on bonds, etc. Right now, the main contingency is death, together with some others like retirement. These decrements are treated stochastically, but interest rates are not.

A major change in the SOA examinations is being contemplated, upon report of a special task force. The general mathematics examinations, i.e., those below Course 140, may be eliminated! Rather, the SOA will rely on colleges to teach these. A general preliminary, or "attractor" examination, stressing business applications of mathematics, may substitute, as well as a Property/Casualty examination and one on statistical modelling. Also, in the upper examinations, jurisdiction-specific "compliance" matters (e.g., what Ontario tax law section covers problem X?) may disappear, although *conceptual* questions may stay/appear. The general idea is "*science, not compliance*". This will eliminate many of the later FSA Courses. Of course, the new stress on conceptual understanding may complicate study and grading. It is easier to grade rote replies than conceptual ones with many "right" answers.

Another emphasis in the reforms is on more and earlier application to real business, not just "theoretical," problems. Education may be restructured into (1) Preliminary (learned in college), (2) Basic (the remaining 100- and present 200-Courses, (3) Advanced (the 300- and the remaining 400-Courses), where one specializes in a "track," and (4) Professional Development (jurisdiction-specific material like tax laws, to be learned outside the SOA exams by continuing, and likely required, education).

A major spur to all this seems to be the decline in traditional employment and an increase in new employment where conceptualizing is important. It must be stressed that the reforms just discussed have not been finalized, and much could change as the SOA membership gives its input, but it shows the general direction the Society may follow.

So much for the examinations. Some recent acts by the SOA aim to bolster ties to colleges and universities. First, the SOA has begun awarding grants to those admitted to doctoral candidacy with a thesis related to actuarial science. The amount is $10,000 per year renewable at most thrice. Second, the SOA awards $5,000 to a college when a full-time faculty member achieves the ASA and $7,500 for the FSA, to promote actuarial education. Third, full-time faculty who belong to the SOA have certain registration and travel fees at SOA meetings waived or paid.

4. RESEARCH. Undergraduate major programs in actuarial science, defined as studies through Course 150, are given in some 60 colleges and universities in North America, with 200 other colleges giving some training. The programs tend to be housed with mathematics or business, with most in mathematics. The "top" research institutions—Harvard, Princeton, Stanford, Berkeley, Caltech, MIT, Chicago, etc.—lack actuarial programs. However, the next "rung" does feature them, as exemplified by most of the Big Ten. A few institutions have graduate studies.

Historically, many actuarial programs were started or encouraged by insurance companies wanting a steady stream of new hires with good math backgrounds and perhaps one or two exams passed. The stress was on teaching in a technical way, rather than research. This hindered achieving tenure for faculty in the larger schools, where research was required. Also, research grants were few, and the small number of actuaries within a given department allowed little chance for face-to-face collaboration. Most college instructors lack the combination of Fellowship in an actuarial society and the PhD, although the number is rising. Among the teaching-oriented institutions, an actuarial program may be viewed favorably as it attracts bright students who necessarily take, and enliven, many regular math and/or business classes.

Insurance companies tend to be conservative about whom they hire. Pathbreaking researchers are generally not desired, except perhaps by one or two very large companies. Most actuaries find academic research of little use in their struggle to complete their usual work. Any academic journals received usually sit on the shelf. This may change as "practical" work grows more complex. The professional societies (the SOA, the CAS, etc.) could help by giving grants for *making research accessible*, rather than merely for the research itself. Right now, academic researchers seem to write mostly for one another. Recall also that actuarial examinations are practical, not research-oriented.

Before giving current research topics, I address related matters. The primary actuarial research journal in North America is the *Transactions of the Society of Actuaries* (TSA), published by the SOA and sent to all members. Printed once a year in hard cover, TSA runs several hundred pages; preprints come on request but only 10% request them. So the harvest is tiny next to all North American math journals! Also, some TSA papers are really data-gathering, such as for mortality rates, and are not what mathematicians would deem "research".

Until recently only SOA members or those sponsored by them could write for the TSA. A paper went to several reviewers who wrote reports. Based on these,

the main reviewer recommended acceptance, conditional acceptance, conditional rejection, or rejection. Some 50%–65% of papers were finally accepted. Short notes and papers on education/exams/pedagogy were lacking; an exception was the address of the incoming SOA President, which dealt with various issues. Referees were not given authors' names ("anonymous refereeing"). An appeals process existed for unhappy submitters.

The SOA Board of Governors has approved new operating procedures and format for the TSA. Examples: open the authorship (i.e., don't limit it to the SOA), make the TSA more of a scientific journal, modify the review process so the editorial board acts more like a collection of associate editors farming out papers to experts where appropriate, train referees, have outside expert referees and perhaps editors, recruit experts to write papers, make the appeals process more thorough, set up a papers contest, recruit papers on general finance, accept papers of four general kinds: current research, actuarial education, comprehensive reviews, and useful insights on "practical" topics. The driving idea: build a body of general financial knowledge for actuaries; i.e., build skills outside pure reckoning of life insurance. Another proposal is to start a second, quarterly journal to ferret out papers of current research interest, to resemble, e.g., the *Journal of Risk and Insurance*, or the *Journal of Finance*. Perhaps some of this will make research more accessible.

Other actuarial journals exist abroad; a big area is Scandinavia, where *The Scandinavian Actuarial Journal* is prominent. Also, in North America there is ARCH (*Journal of the Actuarial Research Clearing House*), published a few times a year based on volume; here authors appear quickly, verbatim and unrefereed. In the U.S. a new *Journal of Actuarial Practice*, where less "theoretical" material may be featured, has begun publication. *Insurance: Mathematics and Economics* is a journal in a related field(s). Also, several newsletters sometimes print shorter "semi-research": *The Actuary* of the SOA and newsletters of SOA sections are examples. A few other bodies such as the American Academy of Actuaries have non-technical magazines. Finally, with the growing link to general finance, "non-actuarial" journals feature articles of interest to research actuaries, and some actuaries write for such journals.

Recent SOA efforts to promote research deserve mention. I have mentioned improving ties to colleges and universities (e.g., PhD grants) and giving 30 exam credits for a research paper. Also, for some years the SOA has had an Actuarial Education and Research Fund (AERF), which sponsors an annual individual grants competition. SOA research moneys have risen from $180,000 in 1986 to about $600,000 in 1995. And the SOA Committee on Knowledge Extension Research has begun a yearly grants competition seemingly like AERF's, but practitioners as well as academics are urged to apply; individual awards do not exceed $10,000. Also, an Actuarial Foundation is being created to help support such efforts. Finally, there is an annual Actuarial Research Conference attended by a few score researchers, giving one of the few good opportunities for face-to-face discussions and in marked contrast to the much larger, less esoteric meetings for practicing actuaries. The number of actuarial researchers is few.

Actuaries use their statistics and mathematics to solve problems in insurance and pensions. In older days, "first generation actuaries" used mostly deterministic models. With stochastic modelling came the "second generation"; a good example is [1] used now for Courses 150 and 151, where contingent events are modelled stochastically and where it is proposed to begin stochastic modelling of interest rates, too. Modern actuaries using stochastic modelling everywhere are "third

generation"; here general finance theory enters, with multivariate stochastic approaches increasingly used.

Here are some important current research topics:

(1) Ordering of risks; i.e., grading risks from weakest to strongest. Probability theory and properties of monotone and convex functions are used. For its mathematical technicality, this area appeals to academics. Applications are wide: premium calculations, optimal reinsurance (see definition in Section 2), ruin probability and portfolio modelling ("portfolio" is the kinds and amounts of insurance products the company offers, and is used to project claims; see (2)). Various kinds of ordering are used: (a) moment order, (b) stochastic order, and (c) "stop-loss" order (risk X is less than risk Y if (i) $E(X^n) \leq E(Y^n)$ for all natural numbers n, (ii) $F_X(t) \geq F_Y(t)$ for all t, (iii) $E[(X-d)_+] \leq E[(Y-d)_+]$ for all d, where $+$ means taking the integral only to the right of d); applications are in optimum premium and benefit calculations and most recently in multistate modelling. Martingale theory also applies here, as well as in calculation of premiums.

(2) Insurance portfolio models. This is very popular with practitioners and academics. The main challenge is to model numbers and amounts of claims (respectively, "frequency" and "severity") in order to calculate total distributions of claims. Many papers exist on both individual and collective risk models, and the area stands actively researched. The most important mathematical tools are probability generating functions, Laplace transforms, and the properties of differential and difference equations.

(3) Credibility; i.e., how much to rely on data A vs. B in predicting the future, where data A come often from your own company and data B from the whole industry (you wish to rely on data A as much as is *credible*). It is popular throughout the profession. General regression, nonlinear regression, Bayesian statistics, and Kalman filtering methods using stochastic differential equations are used and, lately, estimating function techniques from functional analysis. Mathematical statistics and estimation theory enter widely for better results.

(4) Graduation; i.e., smoothing data via math techniques (e.g., we don't want a death rate of .003 at age 35, .009 at age 36, .001 at age 37, .01 at 38). "First generation" actuaries used a deterministic model from numerical analysis. Recent efforts use parametric models (with estimation of parameters), nonlinear regression, spline fitting, "Bayesian" graduation—all from statistics.

(5) Demography. Again, the "first generation" was deterministic; calculus and linear algebra sufficed. But modern actuaries need stochastic life and survival functions, survival models, etc. Advanced probability tools exist but are still foreign to most actuaries.

(6) Interest rate modelling. Again, the first approach was deterministic. Lately, advanced statistical tools such as stochastic differential equations and stochastic integration have come to the fore. The interest rate is often the most important variable in modelling insurance.

(7) Health insurance. In this now-vital area, credibility theory is expected to be fruitful. Fuzzy set theory is also used.

(8) Dynamic solvency testing; i.e., continuous testing to see if the company stays solvent under many scenarios. Stochastic methods enter here, too, as well as computer simulation models.

It is very difficult to identify explicitly all the mathematics and statistics that actuaries use in current research, but this discussion should give the flavor. The

ever-increasing needs of actuarial applications will ever-increasingly require quantitative tools.

5. "PRACTICAL" ISSUES. The supply/demand of jobs has worsened for job-seekers. More people have taken exams, while "downsizing" has hit insurance, too. Top college graduates with some exams passed can still land jobs, but others face hardship. Also, veteran actuaries have lost jobs, a phenomenon almost unknown before. That professional "job-raters" continually rate actuary as the best of jobs ups the supply of job-seekers but not the demand for their services. Tied to all this, "old" actuaries struggle with mathematical and technical advances and are burdened by time pressures as industry economizes. Experience seems worth less when "whiz kids" are worth more, and the mentor-novice bond frays: it saves money to replace veterans earning \$70,000–\$100,000 with youngsters at \$40,000.

Regulatory pressures have increased. Insurers undergo state regulation and are rated by rating agencies (A. M. Best, Standard and Poor's, Moody's, etc.). Standards are stricter following some sensational bankruptcies, and it falls on the actuary to show the company is sound. Accounting regulations imposed by the Financial Accounting Standards Board have multiplied. For example, mutual insurance companies, formerly exempt, must now follow Generally Accepted Accounting Principles.

More knowledge is required of actuaries. General finance is important as actuaries must know derivative securities, general investment strategies, and the like. Demands also grow in reserving/valuation (see Section 2). Formerly, it sufficed to meet legal minimums, but now the valuation actuary must test for solvency under many scenarios and must turn more to stochastic modelling. Also, with more customer awareness and company rivalry, one must design new, saleable products (health insurance?).

Inaugural speeches by incoming SOA Presidents list other concerns: if exams lessen (see Section 3), dues may rise since exam fees are a large part of SOA income, and members may balk; the Canadians, a fifth of the SOA, may slowly secede; six actuarial organizations in North America are too many; general actuarial principles need more development; world ties must be pressed; actuaries must engage more in professional matters (where were they in the health care debates?); continuing education should be required, as it is now for lawyers and CPAs, despite "all those exams"; we must seek a wider base including finance and liberal arts folk with a broad view—but better math skills are also needed; actuaries are too self-effacing and isolated—they must communicate better and cultivate the media (an SOA group is now set up, plus "Actuaries Online", an electronic network among actuaries themselves); actuaries and employers must stop wanting simple point-estimates of risk.

What to tell an actuarial prospect? You must be good, work hard, be broad, communicate well, keep computer-literate, and know a lot of mathematics and statistics. You enter a proud profession. After all, its diligence helped avoid a Savings-and-Loan-type mess. But things are quickly changing. Actuaries are now on Wall Street, plan manufacturing processes, do appraisal work—more than just insurance.

Some advice to the SOA: (1) work toward a full-time President and Board who serve four or more years together, and (2) work with industry so authors get "sabbatical leaves" to come to headquarters, interact, and improve writing skills. Professionalize the profession!

At the end, "actuary" may still be the top occupation, at least after "those awful exams!"

ACKNOWLEDGMENTS. Thanks go to those who helped prepare this paper, especially Harry Panjer and his graduate student Hoque Sharif, whose comments on recent research are incorporated, and Robert Brown for his remarks on recent examination reforms; all are from the University of Waterloo, Ontario. I also thank my wife, Elaine, for proofing and typing the manuscript.

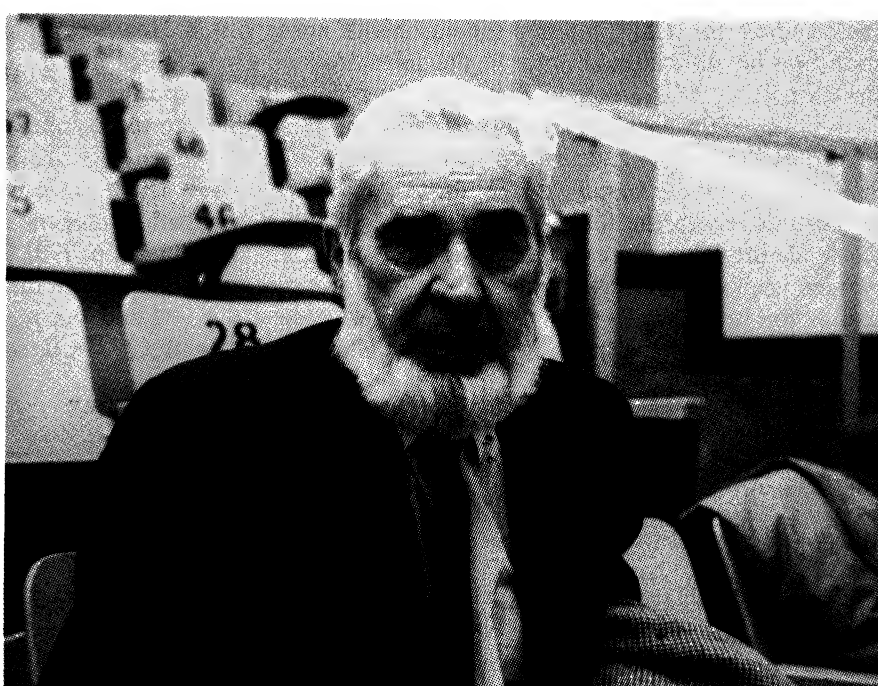
REFERENCES

1. N. L. Bowers, Jr., et al., *Actuarial Mathematics*, Society of Actuaries, 1986.
2. L. M. Friedler, book review of Actuarial Mathematics, *Amer. Math. Monthly* 93 (1986), 489–491.
3. B. Warren, Actuarial science, *Amer. Math. Monthly* 89 (1982), 492–494.

Department of Mathematics & Computer Science
Butler University
Indianapolis, IN 46208
dminassian@butler.edu

PICTURE PUZZLE

(from the collection of Paul Halmos)



When he was Secretary, he was usually smooth-shaven.

(see page 577)

4. CONCLUDING CHALLENGES. Here are some problems you are invited to contemplate. One is an open problem.

1) Can you modify the proof of Theorem 1 so as to allow the series to tend to infinity for various $p \in V_{\mathbb{Q}}$ and to diverge (because the limit of the sequence of partial sums does not exist) for other various $p \in V_{\mathbb{Q}}$?

2) Can you modify the formal power series $F(X)$ in Theorem 2 so that for each $p \in V_{\mathbb{Q}}$, the series converges to a transcendental number in $\mathbb{Q}_p$ when evaluated at any nonzero *algebraic* number $\alpha \in \mathbb{Q}_p$?

3) Can you find a formal power series $F(X) \in \mathbb{Q}[[X]]$ so that for each $p \in V_{\mathbb{Q}}$ and $\alpha \in \mathbb{Q} \setminus \{0\}$, $F(\alpha)$ is transcendental in $\mathbb{Q}_p$ with $F(1) = e$ in $\mathbb{Q}_\infty = \mathbb{R}$?

REFERENCES

1. N. Koblitz, *p-adic Numbers, p-adic Analysis, and Zeta-Functions* (Second Edition), Graduate Texts in Mathematics 58, Springer-Verlag, New York, 1984.
2. A. Ostrowski, Untersuchungen zur arithmetischen Theorie der Körper, *Math. Zeit.* 39, 269–404, 1935.
3. W. M. Schmidt, *Diophantine Equations and Approximation*, Springer Lecture Notes in Mathematics 1467, 1991.

Edward B. Burger
Department of Mathematics
Williams College
Williamstown, MA 01267
Edward.B.Burger@williams.edu

Thomas Struppeck
Zurich / Centre Reinsurance
Zurich, Switzerland

Answer to sequence question on page 538

These are the Catalan numbers, $1, 2, 5, 14, 42, 132, 429, \dots$ *in Catalan*, the language of eastern Spain (Barcelona, Valencia, etc.), Andorra, and nearby regions. They are generated by the recurrence $c(n+1) = c(0)c(n) + c(1)c(n-1) + \cdots + c(n)c(0)$ (with $c(0) = 1$), have the generating function $f(x) = (1 - \sqrt{1-4x})/2x$, and can be written explicitly as $c(n) = (2n)!/(n!(n+1)!)$. They occur in an amazing variety of contexts: Dissection of polygons into triangles, random walk from $(0,0)$ to (n,n) staying above $x = y$, numbers of binary trees, number of ways $2n$ people can shake hands across a circular table without intersecting, ...

Contributed by Solomon W. Golomb
University of Southern California

Answer to Picture Puzzle

(Page 561)

Ed Begle

On the Cyclotomic Polynomial $\Phi_{pq}(X)$

T. Y. Lam and K. H. Leung

The mth cyclotomic polynomial $\Phi_m(X)$ is defined to be $\Pi(X - \zeta)$, where ζ ranges over the primitive mth roots of unity in $\mathbb{C}$. It is well-known that $\Phi_m(X)$ is an irreducible polynomial in $\mathbb{Z}[X]$ with degree $\varphi(m)$, where φ denotes Euler's totient function. In particular, $\Phi_m(X)$ is the minimal polynomial of ζ over $\mathbb{Q}$.

If m_0 denotes the largest square-free factor of m, then $\Phi_m(X) = \Phi_{m_0}(X^{m/m_0})$. Therefore, the computation of $\Phi_m(X)$ can be reduced to the case when $m = pq \cdots$, where $p, q, \ldots$ are distinct primes. Of course, $\Phi_p(X) = X^{p-1} + \cdots + X + 1$, so the next interesting case is when $m = pq$. Here are two explicit examples:

$$\Phi_{15}(X) = X^8 - X^7 + X^5 - X^4 + X^3 - X + 1, \tag{1}$$

$$\Phi_{21}(X) = X^{12} - X^{11} + X^9 - X^8 + X^6 - X^4 + X^3 - X + 1. \tag{2}$$

Let $\Phi_{pq}(X) = \Sigma a_k X^k$. In 1883, Migotti [**Mi**] showed that all $a_k \in \{0, \pm 1\}$. In two earlier articles in this *Monthly*, Beiter [$\mathbf{B_1}$] gave a criterion on k for a_k to be 0, 1 or -1, and Carlitz [**Ca**] computed the number of nonzero a_k's. However, Beiter's criterion is a bit difficult to apply. In the following, we give a quick and natural construction of $\Phi_{pq}(X)$ which, in particular, determines all the coefficients a_k in a simple and explicit way. The method and results in this paper are to be compared with those of Beiter [$\mathbf{B_1}$] and Carlitz [**Ca**].

Our construction is based on the fact that $\varphi(pq) = (p-1)(q-1)$ can be expressed uniquely in the form $rp + sq$ where r, s are non-negative integers. For the existence of such a representation, see [**LeV**, p. 22, Ex. 4] and [**Mo**: p. 799]. The uniqueness is an easy exercise (and will be clear from our arguments). Note that in the expression $(p-1)(q-1) = rp + sq$, we must have $r \le q - 2$ and $s \le p - 2$.

Let ζ be any primitive pqth root of unity. Then $\Phi_q(\zeta^p) = \Phi_p(\zeta^p) = 0$, so by transposition

$$\sum_{i=0}^{r} (\zeta^p)^i = -\sum_{i=r+1}^{q-1} (\zeta^p)^i, \qquad \sum_{j=0}^{s} (\zeta^q)^j = -\sum_{j=s+1}^{p-1} (\zeta^q)^j.$$

Multiplying these and transposing, we get

$$\left(\sum_{i=0}^{r} \zeta^{ip}\right)\left(\sum_{j=0}^{s} \zeta^{jq}\right) - \left(\sum_{i=r+1}^{q-1} \zeta^{ip}\right)\left(\sum_{j=s+1}^{p-1} \zeta^{jq}\right) = 0.$$

Thus, ζ is a zero of the function

$$f(X) := \left(\sum_{i=0}^{r} X^{ip}\right)\left(\sum_{j=0}^{s} X^{jq}\right) - \left(\sum_{i=r+1}^{q-1} X^{ip}\right)\left(\sum_{j=s+1}^{p-1} X^{jq}\right)X^{-pq}. \tag{3}$$

The first product in (3) is clearly a monic polynomial of degree $(p-1)(q-1)$. The second product is a (also monic) polynomial of degree $(p-1)(q-1)-1$ since its lowest term has degree $(r+1)p+(s+1)q-pq=1$, and its highest term has degree $(q-1)p+(p-1)q-pq=(p-1)(q-1)-1$. Therefore, $f(X) \in \mathbb{Z}[X]$ is monic of degree $(p-1)(q-1)=\varphi(pq)$. Since $f(X)$ vanishes on *every* primitive pqth root of unity, we see that $f(X)=\Phi_{pq}(X)$. Next, note that, upon expanding the products in (3), the resulting monomial terms are all different. In fact, if $i, i' \in [0, q-1]$, $j, j' \in [0, p-1]$, and $ip+jq$ is equal to $i'p+j'q$ or to $i'p+j'q-pq$, then $q|(i-i')$. Hence $i=i'$, and similarly $j=j'$. Therefore, we have proved:

Theorem. *The cyclotomic polynomial* $\Phi_{pq}(X)=\Sigma a_k X^k$ *is given by* $f(X)$ *in* (3). *For* $0 \le k \le (p-1)(q-1)$, *we have* (A) $a_k=1$ *if and only if* $k=ip+jq$ *for some* $i \in [0, r]$ *and* $j \in [0, s]$; (B) $a_k=-1$ *if and only if* $k+pq=ip+jq$ *for some* $i \in [r+1, q-1]$ *and* $j \in [s+1, p-1]$; *and* (C) $a_k=0$ *otherwise. The numbers of terms of the former two kinds are, respectively,* $(r+1)(s+1)$ *and* $(p-s-1)(q-r-1)$, *with difference* 1 (*by an explicit computation*).

Note that our proof of this theorem did not use the Möbius-Dedekind formula for $\Phi_m(X)$, which in this case states that

$$\Phi_{pq}(X)=\frac{(X^{pq}-1)(X-1)}{(X^p-1)(X^q-1)}=\frac{\Phi_p(X^q)}{\Phi_p(X)}. \tag{4}$$

From this formula, we see that $\Phi_{pq}(1)=\Sigma_k a_k=1$, which, of course, confirms the last conclusion in the theorem.

As a simple application of the theorem, let us calculate the "middle coefficient" of $\Phi_{pq}(X)$. It is not difficult to check that the following is equivalent to Beiter's Theorem 2 in [**B₁**], although, curiously enough, Beiter stated her result only for the case in which p and q are odd primes.

Corollary. *Assume that* $q>p$, *and let* $l=(p-1)(q-1)/2$. *Then the middle coefficient* a_l *of* $\Phi_{pq}(X)$ *is* $(-1)^r$.

Proof: First assume that $p>2$. Since $(p-1)(q-1)=rp+sq$, r and s have the same parity. If r is *even*, then by (A) of the theorem, $l=(r/2)p+(s/2)q$ shows that $a_l=1$. If r is *odd*, then so is s, and we can write

$$l+pq=\left(\frac{r+q}{2}\right)p+\left(\frac{s+p}{2}\right)q.$$

Since $r \le q-2$ and $s \le p-2$, we have $(r+q)/2 \in [r+1, q-1]$, and $(s+p)/2 \in [s+1, p-1]$. Therefore, (B) of the theorem shows that $a_l=-1$. Finally, let $p=2$. In this case, we have $r=(q-1)/2$, $s=0$, and

$$\Phi_{pq}(X)=\Phi_{2q}(X)=\Phi_q(-X)=\sum_{i=0}^{q-1}(-X)^i,$$

so the middle coefficient is $a_{(q-1)/2}=(-1)^{(q-1)/2}=(-1)^r$. □

Using the explicit form of $f(X)$ in (3), it is also possible to account for the distribution of ± 1's among the nonzero coefficients a_k of $\Phi_{pq}(X)$. We will not go into the details here since this was already done by Carlitz (see the last paragraph of [**Ca**]). Most noteworthy is the consequence that, if we drop the zero coefficients

in $\Phi_{pq}(X)$, the positive and negative terms occur *alternately*, as in the examples (1) and (2). In these two examples, pairs of "alternating" monomial terms of degree $\geq (p-1)(q-1)/2$ occur with the simple pattern $X^{i+1} - X^i$. Indeed, from (4), it is easy to see that this is always the pattern when $p \leq 3$. However, if $p, q > 3$, this pattern no longer persists. For instance, up to (and including) the middle coefficient,[1]

$$\Phi_{55}(X) = X^{40} - X^{39} + X^{35} - X^{34} + X^{30} - X^{28} + X^{25} - X^{23} + X^{20} - \cdots,$$

with gaps of 2 in the exponents in some pairs of alternating terms. The fact that all exponents in the positive terms here are multiples of 5 is not an accident. In fact, if q happens to be of the form $1 + ap$, then $r = a(p-1)$, $s = 0$, and (3) shows that the positive part of $\Phi_{pq}(X)$ is precisely $\sum_{i=0}^{a(p-1)} X^{ip}$.

For related literature on $\Phi_{pq}(X)$, see **[R]**, **[deB]**. For information on $\Phi_{pqt}(X)$ for three distinct primes, see, for instance **[Mi]**, **[Le]**, **[$\mathbf{B_2}$]**, and **[Bm]**.

Note added September 7, 1995. After this note was accepted for publication, we discovered that H. Lenstra had used a similar idea to compute $\Phi_{pq}(X)$ in his report *Vanishing sums of roots of unity*, in *Proc. Bicentennial Congress Wiskundig Genootschap* (Vrije Univ. Amsterdam, 1978), Part II, pp. 249–268. In comparing our approach with Lenstra's, note that the number $r + 1$ (respectively, $s + 1$) in this paper is exactly the inverse of p modulo q (respectively, the inverse of q modulo p).

REFERENCES

[B_1] M. Beiter, The midterm coefficient of the cyclotomic polynomial $F_{pq}(x)$, *Amer. Math. Monthly* 71 (1964), 769–770.

[B_2] M. Beiter, Magnitude of the coefficients of the cyclotomic polynomial F_{pqr}, *Amer. Math. Monthly* 75 (1968), 370–372.

[Bm] D. M. Bloom, On the coefficients of the cyclotomic polynomials, *Amer. Math. Monthly* 75 (1968), 372–377.

[deB] N. G. deBruijn, On the factorization of cyclic groups, *Indag. Math.* 15 (1953), 370–377.

[Ca] L. Carlitz, The number of terms in the cyclotomic polynomial $F_{pq}(x)$, *Amer. Math. Monthly* 73 (1966), 979–981.

[Le] E. Lehmer, On the magnitude of coefficients of the cyclotomic polynomials, *Bull. Amer. Math. Soc.* 42 (1936), 389–392.

[LeV] W. LeVeque, *Topics in Number Theory*, Vol. 1, Addison-Wesley, Reading, Mass., 1956.

[Mi] A. Migotti, Zur Theorie der Kreisteilungsgleichung, *S.-B. der Math.-Naturwiss. Classe der Kaiser. Akad. der Wiss., Wien* 87 (1883), 7–14.

[Mo] Solution to Elementary Problem E1637, *Amer. Math. Monthly* 71 (1964), 799.

[R] L. Rédei, Ein Beitrag zum Problem der Faktorisation von Abelschen Gruppen, *Acta Math. Acad. Sci. Hungar.* 1 (1950), 197–207.

Department of Mathematics
University of California
Berkeley, CA 94720
lam@math.berkeley.edu

Department of Mathematics
National University of Singapore
Singapore 0511
matlkh@leonis.nus.sg

[1]The rest of the coefficients are determined by the fact that $\Phi_m(X)$ is always a self-reciprocal polynomial.

Does $\sum_{n=0}^{\infty} \frac{1}{n!}$ Really Converge? Infinite Series and p-adic Analysis

or

"You can sum some of the series some of the time and some of the series none of the time . . . but can you sum some of the series all of the time?"

Edward B. Burger and Thomas Struppeck

INTRODUCTION. We begin by recalling a fact that most calculus students have memorized, namely,

$$\sum_{n=0}^{\infty} \frac{1}{n!} = e. \tag{1.1}$$

But what exactly does this mean? Convergence is really a topological issue. In particular, if we write $\mathbb{Q}$ for the field of rational numbers and $|\ |$ for the usual absolute value on $\mathbb{Q}$, then $\mathbb{R}$ is the topological completion of $\mathbb{Q}$ with respect to $|\ |$. That is, the set of real numbers $\mathbb{R}$ is the smallest field containing $\mathbb{Q}$ for which all Cauchy sequences of rationals converge. The sequence of partial sums associated with the infinite series (1.1) is a Cauchy sequence of rationals and thus we say the series *converges* in $\mathbb{R}$ and by convention we call the *sum of the series* (the limit of the sequence of partial sums) 'e.' Thus the notion of convergence of (1.1) depends heavily upon the absolute value $|\ |$ on $\mathbb{Q}$. The purpose of this paper is to explore and exploit this theme and to construct some unusual infinite series having some amazing convergence properties.

We are very comfortable with the 'natural' absolute value on $\mathbb{Q}$, but what if we were to consider another absolute value on $\mathbb{Q}$? In fact, are there any other absolute values on $\mathbb{Q}$? What others could there be? For a fixed prime number p, we define the map $|\ |_p$: $\mathbb{Q} \to [0, \infty)$ as follows: we declare $|0|_p = 0$. If $r/s \in \mathbb{Q}$, $r/s \neq 0$, then we may factor it uniquely as

$$\frac{r}{s} = \pm p^{\alpha} p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_L^{\alpha_L},$$

where $p, p_1, \dots, p_L$ are distinct primes, $\alpha \in \mathbb{Z}$, $\alpha_l \in \mathbb{Z}$, $\alpha_l \neq 0$ (as usual, $\mathbb{Z}$ denotes the set of integers). We now define

$$\left|\frac{r}{s}\right|_p = p^{-\alpha}.$$

Example. For $140/297 = 2^2 \cdot 3^{-3} \cdot 5 \cdot 7 \cdot 11^{-1}$, we have

$$\left|\frac{140}{297}\right|_2 = \frac{1}{4}, \quad \left|\frac{140}{297}\right|_3 = 27, \quad \left|\frac{140}{297}\right|_7 = \frac{1}{7}, \quad \left|\frac{140}{297}\right|_{11} = 11, \quad \left|\frac{140}{297}\right|_{17} = 1.$$

It turns out that the map $|\ |_p$ satisfies the conditions required to be an absolute value. In particular,

(*i*) $|x|_p = 0$ if and only if $x = 0$.
(*ii*) $|xy|_p = |x|_p|y|_p$ for all $x, y \in \mathbb{Q}$.
(*iii*) $|x + y|_p \leq |x|_p + |y|_p$ for all $x, y \in \mathbb{Q}$.

The first two properties are easily verified; however, the last property, known as the *triangle inequality*, is not as immediate. To verify (*iii*), we use the observation that if N is an *integer* then $|N|_p \leq 1$. Now let r/s and t/u be two nonzero rational numbers and express them as $p^n r'/s'$ and $p^m t'/u'$, respectively, where $n, m \in \mathbb{Z}$ and p is relatively prime to r', s', t' and u'. Without loss of generality we may assume that $n \leq m$ so $|r/s|_p = p^{-n} \geq |t/u|_p = p^{-m}$. Thus we have

$$\begin{aligned}
\left|\frac{r}{s} + \frac{t}{u}\right|_p &= \left|p^n\left(\frac{r'}{s'} + p^{m-n}\frac{t'}{u'}\right)\right|_p \\
&= p^{-n}\left|\frac{u'r' + p^{m-n}t's'}{s'u'}\right|_p \\
&= p^{-n}\left|(s'u')^{-1}\right|_p|u'r' + p^{m-n}t's'|_p \\
&= p^{-n}|u'r' + p^{m-n}t's'|_p \text{ (note that } u'r' + p^{m-n}t's' \text{ is an integer)} \\
&\leq p^{-n} = \left|\frac{r}{s}\right|_p = \max\left\{\left|\frac{r}{s}\right|_p, \left|\frac{t}{u}\right|_p\right\} \leq \left|\frac{r}{s}\right|_p + \left|\frac{t}{u}\right|_p,
\end{aligned}$$

which establishes the triangle inequality. The absolute value $|\ |_p$ on $\mathbb{Q}$ is called the *p-adic absolute value* (see [1] for a detailed treatment of p-adic analysis). The p-adic absolute value leads to a somewhat strange metric (that is, a strange measure of size) on $\mathbb{Q}$. For example, 2-adically, 20 is very small ($|20|_2 = 1/4$) while $1/24$ is 2-adically huge ($|1/24|_2 = 8$). Basically, the p-adic absolute value of x measures the arithmetical 'size' of x with respect to p in the sense that x is p-adically small whenever a power of p divides the numerator of x. The p-adic absolute value is dramatically different from the usual absolute value on $\mathbb{Q}$; in fact, we note that our proof of the triangle inequality actually proves much more! In particular, we have

(*iii*)′ $|x + y|_p \leq \max\{|x|_p, |y|_p\}$ for all $x, y \in \mathbb{Q}$.

This is plainly stronger than the triangle inequality and so we refer to property (*iii*)′ as the *strong triangle inequality*.

Thus, we see that there are infinitely many different absolute values on $\mathbb{Q}$: the usual one and the p-adic one, for each prime p. Are there others? The answer is no, not really. Suppose that $\|\ \|_1$ and $\|\ \|_2$ are two absolute values on $\mathbb{Q}$. We say that $\|\ \|_1$ and $\|\ \|_2$ are *equivalent* if $\{x \in \mathbb{Q}\colon \|x_1\| \leq 1\} = \{x \in \mathbb{Q}\colon \|x\|_2 \leq 1\}$. In other words, they are equivalent if they induce the same metric topology (see [1]).

Given this, Ostrowski [2] in 1935, proved that any nontrivial absolute value on $\mathbb{Q}$ is equivalent to either the usual absolute value or a p-adic absolute value (the trivial absolute value is one where $|x|_0 = 1$ for all $x \neq 0$). That is, the only (nontrivial) absolute values on the rationals are the arithmetic ones and the usual one. It turns out that all these absolute values work together in the following sense. Suppose that α is a nonzero rational number, $\alpha = \pm p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_L^{\alpha_L}$, where $p_1, \dots, p_L$ are distinct primes, $\alpha_l \in \mathbb{Z}$, $\alpha_l \neq 0$. Then

$$|\alpha| \cdot \prod_{p \text{ prime}} |\alpha|_p = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_L^{\alpha_L} \cdot p_1^{-\alpha_1} p_2^{-\alpha_2} \cdots p_L^{-\alpha_L} = 1. \tag{1.2}$$

We need not concern ourselves with issues of convergence in the infinite product (1.2) since $|\alpha|_p = 1$ for all but finitely many p. The identity (1.2) is known as the *product formula*.

We now write $|\ |_\infty$ for the usual absolute value on $\mathbb{Q}$ and denote the completion $\mathbb{R}$ as $\mathbb{Q}_\infty$. If we allow the possibility that $p = \infty$ and let $V_{\mathbb{Q}} = \{\infty, 2, 3, 5, 7, 11, \dots\}$ ('V' stands for *valuation*, which is another word for absolute value), then the product formula may be expressed in a condensed manner as

$$\prod_{p \in V_{\mathbb{Q}}} |\alpha|_p = 1,$$

for all nonzero rational numbers α. So, in particular, if for a nonzero rational number α, there exists a $\tilde{p}$ such that $|\alpha|_p \leq 1$ for all $p \in V_{\mathbb{Q}} \setminus \{\tilde{p}\}$ with *strict* inequality holding for some of these p's, then the product formula ensures that $|\alpha|_{\tilde{p}} > 1$. That is, in some sense, if α is not zero, then $|\alpha|_p$ cannot be 'small' for all $p \in V_{\mathbb{Q}}$.

For a prime number p, we let $\mathbb{Q}_p$ be the topological completion of $\mathbb{Q}$ with respect to $|\ |_p$. The field $\mathbb{Q}_p$ is called the field of *p-adic numbers*. That is, $\mathbb{Q}_p$ is the smallest field containing $\mathbb{Q}$ such that all Cauchy sequences of rationals converge with respect to $|\ |_p$. How can we think of elements in $\mathbb{Q}_p$? In order to answer this, we turn our attention back to infinite series and observe a dramatic difference between the p-adic absolute value and the usual absolute value. If p is a prime and $\sum_{n=0}^{\infty} a_n$ is an infinite series of rational numbers, then the series converges to a point in $\mathbb{Q}_p$ *if and only if*

$$\lim_{N \to \infty} |a_N|_p = 0. \tag{1.3}$$

This is, of course, a calculus student's dream come true! Note that if the series converges with respect to $|\ |_p$, then for any $\varepsilon > 0$ there exists an integer M_0 so that for all $T > S > M_0$, $|\sum_{n=0}^{T} a_n - \sum_{n=0}^{S} a_n|_p < \varepsilon$. Thus for all $M > M_0$, $|a_m|_p = |\sum_{n=0}^{M} a_n - \sum_{n=0}^{M-1} a_n|_p < \epsilon$, so (1.3) holds. On the other hand, if (1.3) holds, then given $\varepsilon > 0$, there must exist an M_0 so that for all integers $M \geq M_0$, $|a_M|_p < \varepsilon$. Thus by the strong triangle inequality, for any integers $T > S \geq M_0$,

$$\left| \sum_{n=0}^{T} a_n - \sum_{n=0}^{S} a_n \right|_p = \left| \sum_{n=S+1}^{T} a_n \right|_p \leq \max\{|a_{S+1}|_p, |a_{S+2}|_p, \dots, |a_T|_p\} < \varepsilon.$$

Hence the sequence of partial sums is a Cauchy sequence in $\mathbb{Q}_p$ and thus $\sum_{n=0}^{\infty} a_n$ converges in $\mathbb{Q}_p$.

So how can we think of elements in $\mathbb{Q}_p$? We may view them as infinite series! To inspire this notion, let us return momentarily to $\mathbb{R}$. We may express $\alpha \in \mathbb{R}$ in

its decimal expansion as the convergent series

$$\alpha = \sum_{n=l}^{\infty} d_n 10^{-n},$$

where l is an integer and each d_n is an integer, $0 \le d_n \le 9$ (usually we just write the digits in juxtaposition as $\alpha = d_l d_{l+1} d_{l+2} \cdots$, where we place the decimal point between d_0 and d_1). Now returning to the p-adic case, it turns out that every element $\delta \in \mathbb{Q}_p$ has a *p-adic expansion* given by the convergent series

$$\delta = \sum_{n=l}^{\infty} d_n p^n,$$

where l is again an integer and each d_n is an integer, $0 \le d_n \le p - 1$ (see **[1]** for details). This is the p-adic analogue of the decimal expansion. However, note that in the decimal expansion, the exponents of 10 are decreasing while the exponents of p in the p-adic expansion are *increasing*! This appears strange because we are so accustomed to series converging in $\mathbb{R}$. Notice, however, that $|d_n p^n|_p = p^{-n} \to 0$ as $n \to \infty$, so from our previous observation on infinite series, we conclude that the series does converge in $\mathbb{Q}_p$. As an illustration, the 7-adic expansion of $-31/336 \in \mathbb{Q}_7$ is given by

$$-\frac{31}{336} = 3 \cdot 7^{-1} + 4 \cdot 7^0 + 3 \cdot 7^1 + 4 \cdot 7^2 + \cdots$$

(verify this: it's fun!).

Other Examples. $\Sigma_{n=0}^{\infty} 3^n = -1/2$ in $\mathbb{Q}_3$ (it is a convergent geometric series 3-adically), but this series *diverges* in all other $\mathbb{Q}_p$, $p \ne 3$ (note that for primes $p \ne 3, |3^n|_p = 1 \not\to 0$ as $n \to \infty$). Similarly, $\Sigma_{n=0}^{\infty} 1/n!$ converges in $\mathbb{Q}_\infty$ but diverges in all other $\mathbb{Q}_p$, $p \ne \infty$. Finally we note that $\Sigma_{n=1}^{\infty} 1/n$ diverges in $\mathbb{Q}_p$ for *all p*. So we have seen that we can sum some of the series some of the time and some of the series none of the time. This leads us to our first question.

Question 1. Is it possible to produce an infinite series of rational numbers so that for each $p \in V_{\mathbb{Q}}$, the series converges (with respect to $|\ |_p$) in $\mathbb{Q}_p$?

Here is a heuristic argument to show that no such series can exist. For $\Sigma_{n=0}^{\infty} a_n$ to converge in $\mathbb{Q}_p$, p prime, it must be the case that $\lim_{N \to \infty} |a_N|_p = 0$. That is, for all sufficiently large n, the prime p occurs only in the numerators of a_n and to arbitrarily large powers. If the series converged in all $\mathbb{Q}_p$, p prime, then that implies that for all sufficiently large n, all the primes occur in the numerators of a_n. Thus, by the product formula, the series must diverge in $\mathbb{Q}_\infty = \mathbb{R}$. For example, consider $\Sigma_{n=0}^{\infty} n!$. This converges in $\mathbb{Q}_p$ for all primes p, but diverges in $\mathbb{Q}_\infty$. Unfortunately, this heuristic is *not* correct.

We now construct an explicit example of an infinite series of rational numbers that converges with respect to each absolute value of $\mathbb{Q}$. We begin with the series $\Sigma_{n=0}^{\infty} n!$, which converges with respect to each p-adic absolute value. It suffices to modify the series so it converges with respect to the usual absolute value. We do this by considering the series $\Sigma_{n=0}^{\infty} n!/(n!)^2$. Well, this does converge in $\mathbb{Q}_\infty$, but sadly, as we have already observed, it diverges with respect to every p-adic

absolute value. It appears as though we have done more harm than good! But now we modify this series ever so slightly. In particular, we consider the series

$$\sum_{n=0}^{\infty} \frac{n!}{(n!)^2 + 1}. \tag{1.4}$$

One may (and should!) verify that this series converges with respect to every absolute value on $\mathbb{Q}$. This answers our first question.

A natural question now is: what does the series (1.4) sum to in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$? In particular, does there exist a p so that the above sum is algebraic in $\mathbb{Q}_p$? Or is there a p for which the sum is transcendental in $\mathbb{Q}_p$? These questions appear to be very difficult to answer. However, we may ask a more general question.

Question 2. Is it possible to construct an infinite series of rational numbers so that the series converges in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$ and for each $p \in V_{\mathbb{Q}}$, the sum of the series is a rational number?

The answer is yes and, in fact, we will discover that much more is true. In particular, we have the following.

Theorem 1. *For each $p \in V_{\mathbb{Q}}$, let $\alpha_p \in \mathbb{Q}_p$ be given. Then there exists an infinite series of rational numbers $\sum_{n=0}^{\infty} a_n$, with $a_n > 0$ for all $n \geq 1$, such that the series converges to α_p in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$.*

As an immediate consequence we conclude that given any rational number α, if we select $\alpha_p = \alpha$ for each $p \in V_{\mathbb{Q}}$, then there exists an infinite series of rationals such that the series converges to α in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$. In [**1**, p. 85], Koblitz asked if there exists an infinite series of rational numbers converging for all $p \in V_{\mathbb{Q}}$ such that it converges to a rational number for some $p \in V_{\mathbb{Q}}$; Koblitz remarked [**1**, p. 142] that neither an example nor a proof that it is impossible is known. Theorem 1 answers this question. Moreover, in the proof of Theorem 1 we give an algorithm for generating the series.

Of course Theorem 1 enables us to create a variety of series with unusual properties. As an illustration, let $g: V_{\mathbb{Q}} \to \mathbb{Q}$ be a bijection. Then we may construct an infinite series of rational numbers such that for each $p \in V_{\mathbb{Q}}$, the series converges to $g(p)$ in $\mathbb{Q}_p$. That is, given any rational number r/s, there exists a unique $p \in V_{\mathbb{Q}}$ so that the series converges in $\mathbb{Q}_p$ to r/s. Thus there exists an infinite series such that it sums to every rational number (when the series is viewed in the appropriate completion $\mathbb{Q}_p$)!

We construct the series of Theorem 1 by using a Cantor diagonalization argument to introduce a new prime p at each stage while continually getting closer to the appropriate 'targets' for the previous p's. We build the a_n's by using the Chinese remainder theorem for p prime together with a density argument for $p = \infty$. To illustrate this basic theme, we note that the infinite series $\sum_{n=0}^{\infty} 1/n!$ converges to e in $\mathbb{Q}_\infty$ and diverges to infinity in $\mathbb{Q}_p$, for each prime p. Thus, as another amusing example, using the algorithm given in the proof of Theorem 1, we may build an infinite series of rational numbers, $\sum_{n=0}^{\infty} a_n$, so that it converges to e in $\mathbb{Q}_\infty$ and converges to 0 in $\mathbb{Q}_p$ for all primes p. We give below the first few values of a_n together with the first few partial sums (factored into primes). Note that at

each stage a new prime is introduced in the numerator of the partial sum and, once it is introduced, it continues to appear with greater and greater exponents. Thus the partial sums are p-adically approaching zero for all primes p while they approach e in $\mathbb{Q}_\infty$. The Cantor diagonalization ensures that every prime p will make it into the numerator at some point; we just have to be patient!

$$2 + \frac{2}{3} = \frac{2^3}{3} \approx 2.666\cdots$$

$$2 + \frac{2}{3} + \frac{8}{159} = \frac{2^4 3^2}{53} \approx 2.71698113\cdots$$

$$2 + \frac{2}{3} + \frac{8}{159} + \frac{144}{168487} = \frac{2^6 3^3 5}{(11)(17^2)} \approx 2.7178357\cdots$$

$$2 + \frac{2}{3} + \frac{8}{159} + \frac{144}{168487} + \frac{397440}{1060988071} = \frac{2^6 3^4 5^2 7}{(13)(25673)} \approx 2.7182103\cdots$$

$$\begin{aligned} &2 + \frac{2}{3} + \frac{8}{159} + \frac{144}{168487} + \frac{397440}{1060988071} + \frac{17909035200}{257300603913301} \\ &= \frac{2^7 3^5 5^3 7^2 11}{(14723)(52363)} \approx 2.718279995\cdots \end{aligned}$$

$$\begin{aligned} &2 + \frac{2}{3} + \frac{8}{159} + \frac{144}{168487} + \frac{397440}{1060988071} + \frac{17909035200}{257300603913301} \\ &+ \frac{17934123171888000}{178482991548071117739339} \\ &= \frac{2^8 3^6 5^4 7^3 11^2 13}{(1847929)(12528259)} \approx 2.7182809999991\cdots \end{aligned}$$

The Cantor diagonalization technique allows us to escape the previous heuristic argument. In fact, the heuristic argument does show that it is impossible to have uniform convergence in all $\mathbb{Q}_p$, $p \in V_{\mathbb{Q}}$, simultaneously. As will be evident from the proof of Theorem 1, the series may be constructed so that for each $p \in V_{\mathbb{Q}}$, the series converges as fast as desired from some point on (which will depend upon p). That is, for each $p \in V_{\mathbb{Q}}$, we may ensure that as $N \to \infty$, $\left| \alpha_p - \sum_{n=0}^{N} a_n \right|_p \to 0$ as rapidly as we wish. If we now consider the formal power series associated with the series of Theorem 1,

$$F(X) = \sum_{n=0}^{\infty} a_n X^n,$$

then plainly $F(1)$ converges to α_p in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$. By our remark on the rate of convergence, we may build the sequence $\{a_n\}_{n=0}^{\infty}$ so that $F(X)$ is an entire function in $\mathbb{Q}_p$ for each p. Finally, we show that there exist power series $F(X)$ that converge to transcendental numbers in $\mathbb{Q}_p$ for all $p \in V_{\mathbb{Q}}$ and *all* $X \in \mathbb{Q}$, $X \neq 0$. In particular, we produce the following explicit transcendence result.

Theorem 2. *Let $F(X)$ be the formal power series defined by*

$$F(X) = \sum_{n=0}^{\infty} \left(\frac{n!}{(n!)^2 + 1} \right)^{(n!)^3} X^n,$$

and let α be a nonzero rational number. Then for every $p \in V_{\mathbb{Q}}$, $F(\alpha)$ converges to a transcendental number in $\mathbb{Q}_p$.

Our method is to ensure that the coefficients in the power series approach zero so quickly with respect to $| \ |_p$, that $F(\alpha)$ must be transcendental by an application of an important theorem due to Liouville. We describe this theme in Section 3.

Thus the moral of all this is that when we write down an infinite series of rational numbers and claim it converges, we must be careful to state where we are viewing the convergence; otherwise, as we have seen, it may converge to many different sums in different completions. We summarize our moral in the following poem.

If you claim a series sums to S
Your metric you must not suppress
The danger, you can now see
Is that another may disagree
And you both may be right: what a mess!

2. THE CONSTRUCTION OF THE SERIES. For an integer $m \geq 2$, we define the set $\mathbf{U}(m) \subseteq \mathbb{Q}$ by

$$\mathbf{U}(m) = \left\{ \frac{r}{s} : r \equiv s \equiv 1 \bmod m \right\}.$$

We begin with the following elementary lemma which we state without proof.

Lemma 3. *Let $m \geq 2$ be an integer. Then $\mathbf{U}(m)$ is a dense subset of $\mathbb{R}$. In particular, for $\varepsilon > 0$ and $a/b \in \mathbb{Q}$, if n is an integer satisfying $n > (|a-b|_\infty - b\varepsilon)/(b^2 m\varepsilon)$, then*

$$\left| \frac{a}{b} - \frac{anm+1}{bnm+1} \right|_\infty < \varepsilon.$$

Let $\mathscr{F}$ be a finite collection of distinct primes. For each $p \in \mathscr{F}$, let δ_p be a nonzero element of $\mathbb{Q}_p$ and denote its p-adic expansion as

$$\delta_p = \sum_{n=l_p}^{\infty} d(p,n) p^n,$$

where $0 \leq d(p,n) \leq p-1$ for each n and $d(p, l_p) \neq 0$. We note that the strong triangle inequality implies $|\delta_p|_p = p^{-l_p}$. Next we define the rational number $\Upsilon = \Upsilon(\{\delta_p\}_{p \in \mathscr{F}})$ by $\Upsilon = \prod_{p \in \mathscr{F}} p^{l_p}$.

Lemma 4. *Given $\mathscr{F}$, $\{\delta_p\}_{p \in \mathscr{F}}$, and Υ as in the preceding paragraph, there exists an integer $M > 0$ such that*

$$|\delta_p - M\Upsilon|_p < |\delta_p|_p$$

for all primes $p \in \mathscr{F}$.

Proof: For each prime $p \in \mathscr{F}$, we define $\Upsilon_p = \Upsilon p^{-l_p}$. Thus $\Upsilon_p p^{l_p} = \Upsilon$. Next we write $\Upsilon_p = m_p/n_p$, where m_p and n_p are integers relatively prime to p. We now consider the following finite collection of simultaneous linear congruences:

$$m_p x \equiv n_p d(p, l_p) \bmod p, \quad \text{for each } p \in \mathscr{F}. \tag{2.1}$$

By the Chinese remainder theorem, we may find a solution to this system, say $x = M > 0$. Therefore for each $p \in \mathscr{F}$, there exists some integer t_p so that

$$\frac{m_p}{n_p} M = d(p, l_p) + p\frac{t_p}{n_p},$$

which implies

$$M\Upsilon = d(p, l_p)p^{l_p} + p^{l_p+1}\left(\frac{t_p}{n_p}\right).$$

Since $n_p \not\equiv 0 \bmod p$, we conclude that the first term in the p-adic expansion for $M\Upsilon$ is $d(p, l_p)p^{l_p}$, which we recall is the first term in the p-adic representation for δ_p. Hence, $|\delta_p - M\Upsilon|_p < |p^{l_p}|_p = |\delta_p|_p$. □

Corollary 5. *Let $\mathscr{F}$, $\{\delta_p\}_{p \in \mathscr{F}}$, Υ, and M be as in Lemma 4. If $\tilde{p}$ is a prime in $\mathscr{F}$, then for any $u \in \mathbf{U}(\tilde{p})$,*

$$|\delta_{\tilde{p}} - uM\Upsilon|_{\tilde{p}} < |\delta_{\tilde{p}}|_{\tilde{p}}.$$

Proof: If we write $u = r/s$ with $r \equiv s \equiv 1 \bmod \tilde{p}$, then it follows that $|1 - r/s$ $|(s - r)/s|_{\tilde{p}} = |s - r|_{\tilde{p}} \leq \tilde{p}^{-1}$. Thus the strong triangle inequality and Lemr reveal

$$\begin{aligned}
|\delta_{\tilde{p}} - uM\Upsilon|_{\tilde{p}} &= |\delta_{\tilde{p}} - M\Upsilon + M\Upsilon - uM\Upsilon|_{\tilde{p}} \\
&\leq \max\left\{|\delta_{\tilde{p}} - M\Upsilon|_{\tilde{p}}, |M\Upsilon - uM\Upsilon|_{\tilde{p}}\right\} \\
&< \max\left\{|\delta_{\tilde{p}}|_{\tilde{p}}, |M\Upsilon|_{\tilde{p}}\left|1 - \frac{r}{s}\right|_{\tilde{p}}\right\} \\
&\leq \max\left\{|\delta_{\tilde{p}}|_{\tilde{p}}, |\delta_{\tilde{p}}|_{\tilde{p}}\tilde{p}^{-1}\right\} = |\delta_{\tilde{p}}|_{\tilde{p}},
\end{aligned}$$

which establishes the desired inequality. □

Proof of Theorem 1: We will define the rational numbers a_n inductively. We begin by setting $a_0 = [\alpha_\infty - 1]$, where $[x]$ denotes the integer part of x. Next we write $\mathscr{F}_N$ for the set of the first N primes. Thus, for example, $\mathscr{F}_0 = \varnothing$ (the empty set), $\mathscr{F}_1 = \{2\}$ and $\mathscr{F}_2 = \{2, 3\}$. For our inductive hypothesis, we suppose that $a_0, a_1, \ldots, a_N$ have all been defined so that the following three conditions hold:

(*i*) $a_n \in \mathbb{Q}$ for all $0 \leq n \leq N$ and $a_n > 0$ for $n \neq 0$.
(*ii*) If we write $S_N = \sum_{n=0}^{N} a_n$ for the N^{th} partial sum, then $0 < \alpha_\infty - S_N < 2^{-(N-1)}$.
(*iii*) For each prime p in $\mathscr{F}_N$, either $S_{N-1} = \alpha_p$ or $|\alpha_p - S_N|_p < |\alpha_p - S_{N-1}|_p$.

Since $a_0 \in \mathbb{Q}$, $0 < \alpha_\infty - a_0 < 2$ and $\mathscr{F}_0 = \phi$, conditions (*i*), (*ii*), and (*iii*) are all satisfied when $N = 0$.

We now construct the term a_{N+1}. For each prime $p \in \mathscr{F}_{N+1}$, we write $\delta_p = \alpha_p - S_N$ and let $\tilde{\mathscr{F}}_{N+1} = \{p \in \mathscr{F}_{N+1}\colon \delta_p \neq 0\}$. In case $\tilde{\mathscr{F}}_{N+1} = \varnothing$, we declare $M\Upsilon = 1$. Otherwise by Lemma 4, there exists a positive integer M such that

$|\delta_p - M\Upsilon|_p < |\delta_p|_p$ for all primes $p \in \tilde{\mathscr{F}}_{N+1}$. If we write P_{N+1} for the product of the first $N+1$ primes, then by Lemma 3 there exists a rational number $u \in \mathrm{U}(P_{N+1})$ such that

$$\left|\frac{\alpha_\infty - S_N}{M\Upsilon} - u\right|_\infty < (2M\Upsilon)^{-1}|S_N - \alpha_\infty|_\infty, \tag{2.2}$$

and

$$0 < u < \frac{\alpha_\infty - S_N}{M\Upsilon}. \tag{2.3}$$

We note that $\mathrm{U}(P_{N+1}) = \bigcap_{p \in \mathscr{F}_{N+1}} \mathrm{U}(p)$, and therefore by Corollary 5 we have

$$|\delta_p - uM\Upsilon|_p < |\delta_p|_p \tag{2.4}$$

for all primes $p \in \tilde{\mathscr{F}}_{N+1}$. We now define $a_{N+1} = uM\Upsilon$. In view of inequalities (2.2), (2.3), and (2.4), conditions (i), (ii), and (iii) are satisfied for the index $N+1$.

Finally, we show that $\sum_{n=0}^{\infty} a_n$ converges to α_p in $\mathbb{Q}_p$ for each $p \in V_{\mathbb{Q}}$. By condition (ii), the series converges in $\mathbb{Q}_\infty = \mathbb{R}$ to α_∞. Suppose now that p is prime. Since $\{S_n\}_{n=0}^{\infty}$ is a monotonically increasing sequence of rational numbers, $\alpha_p = S_J$ for at most one J. From this along with inequality (2.4), we conclude that there exists an N so that

$$0 < |\alpha_p - S_{n+1}|_p < |\alpha_p - S_n|_p$$

for all $n \geq N$. Thus the sequence $\{|\alpha_p - S_n|_p\}_{n=N}^{\infty}$ is a strictly decreasing infinite sequence of integral powers of p. Therefore, $\lim_{n\to\infty}|\alpha_p - S_n|_p = 0$, and hence the series converges in $\mathbb{Q}_p$ to α_p.

We remark that one may replace the upper bound of $|\delta_p|_p$ in Lemma 4 by any positive real number. This is accomplished by modifying the congruences of (2.1) and selecting suitably large integer powers of p for each modulus. Hence by modifying the value of Υ, the series may be constructed so that at each prime p, the series converges p-adically to α_p faster than any specified series. Similarly, we may replace the upper bound of $2^{-(N-1)}$ in condition (ii) by any other monotonically decreasing positive sequence approaching zero. Thus we may also ensure fast convergence in $\mathbb{Q}_\infty$. Therefore, the rates of convergence may be varied independently for each $p \in V_{\mathbb{Q}}$.

3. REMARKS ON FORMAL POWER SERIES. To illustrate the basic idea, let $p \in V_{\mathbb{Q}}$ be fixed. We begin by building a formal power series in $\mathbb{Q}[[X]]$ that converges to transcendental numbers in $\mathbb{Q}_p$ for all nonzero rational X. We first state an important theorem from diophantine approximation due to Liouville (for further details see [3]).

Liouville's Theorem. *For $p \in V_{\mathbb{Q}}$, let $\alpha \in \mathbb{Q}_p$ be an algebraic number of degree $d \geq 1$ over $\mathbb{Q}$. Then there exists a constant $c = c(\alpha) > 0$ such that for all rational numbers r/s, $\alpha \neq r/s$,*

$$\frac{c}{h(r/s)^d} \leq \left|\alpha - \frac{r}{s}\right|_p,$$

where $h(r/s) = max\{|r|_\infty, |s|_\infty\}$, $\gcd(r, s) = 1$.

Next we select a sequence of rationals $\{\beta_0, \beta_1, \dots\}$ so that

$$0 < |\beta_N|_p \le N^{-N} h_1(\beta_0, \beta_1, \dots, \beta_{N-1})^{-(N-1)^2}, \tag{3.1}$$

where h_1 is defined

$$h_1(\beta_0, \beta_1, \dots, \beta_{N-1}) = \max\{h(\beta_0), h(\beta_1), \dots, h(\beta_{N-1})\}, \tag{3.2}$$

and build the associated formal power series

$$F(X) = \sum_{n=0}^{\infty} \beta_n X^n.$$

Suppose that $\gamma \in \mathbb{Q}$, $\gamma \neq 0$ with $h(\gamma) = M$. We define $Q_N = \sum_{n=0}^{N} \beta_n \gamma^n \in \mathbb{Q}$. It follows from (3.2) that

$$\begin{aligned} h(Q_N) &\le (N+1) h_1(\beta_0, \beta_1, \dots, \beta_N)^{N+1} h(\gamma)^N \\ &= (N+1) h_1(\beta_0, \beta_1, \dots, \beta_N)^{N+1} M^N. \end{aligned} \tag{3.3}$$

Next from (3.1), we observe that for all integers $N > M$,

$$\begin{aligned} |F(\gamma) - Q_N|_p &= \left| \sum_{n=N+1}^{\infty} \beta_n \gamma^n \right|_p \\ &\le \sum_{n=N+1}^{\infty} |\beta_n|_p |\gamma|_p^n \\ &\le \sum_{n=N+1}^{\infty} n^{-n} h_1(\beta_0, \beta_1, \dots, \beta_{n-1})^{-(n-1)^2} M^n \\ &\le \sum_{n=N+1}^{\infty} h_1(\beta_0, \beta_1, \dots, \beta_{n-1})^{-(n-1)^2} \\ &\le \sum_{n=N+1}^{\infty} h_1(\beta_0, \beta_1, \dots, \beta_N)^{-(n-1)^2} \\ &= h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \sum_{n=N+1}^{\infty} h_1(\beta_0, \beta_1, \dots, \beta_N)^{-(n-1)^2+N^2} \\ &\le h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \sum_{n=0}^{\infty} h_1(\beta_0, \beta_1, \dots, \beta_N)^{-n} \\ &\le h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \sum_{n=0}^{\infty} 2^{-n} = 2 h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2}. \end{aligned}$$

Thus we have

$$|F(\gamma) - Q_N|_p \le 2 h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \tag{3.4}$$

for all integers $N > M$.

If we now assume that $F(\gamma)$ is *algebraic* of degree d, then by Liouville's theorem there exists a constant $c = c(F(\gamma)) > 0$ so that

$$\frac{c}{h(Q_N)^d} \le |F(\gamma) - Q_N|_p.$$

Thus, in view of (3.3) and (3.4) the preceding inequality yields

$$
\begin{aligned}
0 < c/2 &\le h(Q_N)^d h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \\
&\le (N+1)^d M^{dN} h_1(\beta_0, \beta_1, \dots, \beta_N)^{d(N+1)} h_1(\beta_0, \beta_1, \dots, \beta_N)^{-N^2} \\
&= \frac{(N+1)^d}{h_1(\beta_0, \beta_1, \dots, \beta_N)^{N^2/3}} \cdot \frac{M^{dN}}{h_1(\beta_0, \beta_1, \dots, \beta_N)^{N^2/3}} \cdot \frac{h_1(\beta_0, \beta_1, \dots, \beta_N)^{d(N+1)}}{h_1(\beta_0, \beta_1, \dots, \beta_N)^{N^2/3}}.
\end{aligned}
$$

However, as N approaches infinity, the preceding expression approaches zero, which is impossible since it is at least $c/2$, which is positive. Therefore, $F(\gamma)$ is not algebraic and hence is a transcendental number in $\mathbb{Q}_p$.

We now construct the coefficients $\{\beta_0, \beta_1, \dots\}$ so that for any given $p \in V_{\mathbb{Q}}$ there exists an integer $N(p)$ so that for all integers $N > N(p)$, (3.1) holds. Thus we produce transcendence in $\mathbb{Q}_p$ for all $p \in V_{\mathbb{Q}}$! Subsequently, we demonstrate that for each p, the coefficients occurring in the power series of Theorem 2 satisfy (3.1) for all sufficiently large N. To do so, it will be useful to have an upper bound for $|n!|_p$, where $n > 0$ is an integer and p is a prime. For a positive integer n and prime p, we write the p-adic expansion of n as

$$n = a_0 + a_1 p + a_2 p^2 + \cdots + a_L p^L,$$

where the a_l's are integers satisfying $0 \le a_l \le p-1$, $a_L \ne 0$. We define the integer $A_p(n)$ by $A_p(n) = a_0 + a_1 + \cdots + a_L$. Our upper bound on $|n!|_p$ is a consequence of the following well-known result.

Lemma 6. *If p be a prime and n a positive integer, then*

$$|n!|_p = p^{-(n - A_p(n))/(p-1)}. \tag{3.5}$$

Proof: We prove this by induction on n. If $n = 1$, then clearly (3.5) holds. We now assume that (3.5) holds for $n = N - 1$. If we write

$$N - 1 = a_0 + a_1 p + a_2 p^2 + \cdots + a_L p^L,$$

then it follows that

$$A_p(N) = \begin{cases} A_p(N-1) + 1 & \text{if } a_0 < p - 1 \\ A_p(N-1) - T(p-1) + 1 & \text{if } a_t = p-1, 0 \le t \le T-1, \\ & \qquad \text{and } a_T \ne p - 1. \end{cases}$$

If $a_0 < p - 1$, then $|N|_p = 1$. Hence, given the value of $A_p(N)$ in this case and our inductive hypothesis, we see that

$$
\begin{aligned}
|N!|_p = |N|_p |(N-1)!|_p &= p^{-(N-1-A_p(N-1))/(p-1)} \\
&= p^{-(N-(A_p(N-1)+1))/(p-1)} = p^{-(N-A_p(N))/(p-1)}.
\end{aligned}
$$

If $a_t = p - 1$ for $0 \le t \le T - 1$ and $a_T \ne p - 1$, then $|N|_p = p^{-T}$. Thus by our inductive hypothesis,

$$
\begin{aligned}
|N!|_p = |N|_p |(N-1)!|_p &= p^{-T} p^{-(N-1-A_p(N-1))/(p-1)} \\
&= p^{-(N-(A_p(N-1)-T(p-1)+1))/(p-1)} = p^{-(N-A_p(N))/(p-1)},
\end{aligned}
$$

which completes the proof. □

Corollary 7. *If p is a prime, then for all sufficiently large integers n,*

$$|n!|_p \le p^{-n/(2p-2)}.$$

Proof: Again we write

$$n = a_0 + a_1 p + a_2 p^2 + \cdots + a_L p^L,$$

where the a_l's are integers satisfying $0 \le a_l \le p - 1$, $a_L \ne 0$. Therefore $n \ge p^L$ and thus $L \le \log n/\log p$. This yields

$$A_p(n) \le \frac{(p-1)}{\log p}\log n + (p-1),$$

and thus for sufficiently large n we have $A_p(n) \le n/2$. In view of Lemma 6 and the previous inequality, we conclude that $|n!|_p \le p^{-n/(2p-2)}$. □

Proof of Theorem 2: For integers $n = 0, 1, \ldots,$ we define

$$\beta_n = \left(\frac{n!}{(n!)^2 + 1}\right)^{(n!)^3}.$$

From our remarks at the beginning of this section, we see that to prove the theorem, it is enough to show that for each $p \in V_{\mathbb{Q}}$, (3.1) holds for all sufficiently large integers N.

If $p = \infty$, we note that

$$|\beta_N|_\infty = \left(\frac{N!}{(N!)^2 + 1}\right)^{(N!)^3} \le (N!)^{-(N!)^3}.$$

On the other hand, for large N we have

$$\begin{aligned} N^{-N} h_1(\beta_0, \beta_1, \ldots, \beta_{N-1})^{-(N-1)^2} &= N^{-N}\left(((N-1)!)^2 + 1\right)^{-(N-1)^2((N-1)!)^3} \\ &\ge (N!)^{-N-(N-1)^2((N-1)!)^3} \\ &\ge (N!)^{-(N!)^3} \ge |\beta_N|_\infty. \end{aligned}$$

Thus the theorem holds for $p = \infty$.

We now assume that p is a prime. In this case, we apply Corollary 7 to deduce that for all sufficiently large N,

$$|\beta_N|_p = \left|\left(\frac{N!}{(N!)^2 + 1}\right)^{(N!)^3}\right|_p = |N!|_p^{(N!)^3} \le p^{-(N!)^3 N/(2p-2)}.$$

As we have already observed,

$$N^{-N} h_1(\beta_0, \beta_1, \ldots, \beta_{N-1})^{-(N-1)^2} \ge (N!)^{-N-(N-1)^2((N-1)!)^3}.$$

For N sufficiently large we note that

$$(N!)^{-N-(N-1)^2((N-1)!)^3} \ge p^{-(N!)^3 N/(2p-2)}$$

and thus for the prime p, (3.1) holds for all sufficiently large N. Hence for any $p \in V_{\mathbb{Q}}$ and any nonzero rational α, we conclude that $F(\alpha)$ is a transcendental number in $\mathbb{Q}_p$. □

4. CONCLUDING CHALLENGES. Here are some problems you are invited to contemplate. One is an open problem.

1) Can you modify the proof of Theorem 1 so as to allow the series to tend to infinity for various $p \in V_{\mathbb{Q}}$ and to diverge (because the limit of the sequence of partial sums does not exist) for other various $p \in V_{\mathbb{Q}}$?

2) Can you modify the formal power series $F(X)$ in Theorem 2 so that for each $p \in V_{\mathbb{Q}}$, the series converges to a transcendental number in $\mathbb{Q}_p$ when evaluated at any nonzero *algebraic* number $\alpha \in \mathbb{Q}_p$?

3) Can you find a formal power series $F(X) \in \mathbb{Q}[[X]]$ so that for each $p \in V_{\mathbb{Q}}$ and $\alpha \in \mathbb{Q} \setminus \{0\}$, $F(\alpha)$ is transcendental in $\mathbb{Q}_p$ with $F(1) = e$ in $\mathbb{Q}_\infty = \mathbb{R}$?

REFERENCES

1. N. Koblitz, *p-adic Numbers, p-adic Analysis, and Zeta-Functions* (Second Edition), Graduate Texts in Mathematics 58, Springer-Verlag, New York, 1984.
2. A. Ostrowski, Untersuchungen zur arithmetischen Theorie der Körper, *Math. Zeit.* 39, 269–404, 1935.
3. W. M. Schmidt, *Diophantine Equations and Approximation*, Springer Lecture Notes in Mathematics 1467, 1991.

Edward B. Burger
Department of Mathematics
Williams College
Williamstown, MA 01267
Edward.B.Burger@williams.edu

Thomas Struppeck
Zurich / Centre Reinsurance
Zurich, Switzerland

Answer to sequence question on page 538

These are the Catalan numbers, $1, 2, 5, 14, 42, 132, 429, \ldots$ *in Catalan*, the language of eastern Spain (Barcelona, Valencia, etc.), Andorra, and nearby regions. They are generated by the recurrence $c(n+1) = c(0)c(n) + c(1)c(n-1) + \cdots + c(n)c(0)$ (with $c(0) = 1$), have the generating function $f(x) = (1 - \sqrt{1-4x})/2x$, and can be written explicitly as $c(n) = (2n)!/(n!(n+1)!)$. They occur in an amazing variety of contexts: Dissection of polygons into triangles, random walk from $(0,0)$ to (n,n) staying above $x = y$, numbers of binary trees, number of ways $2n$ people can shake hands across a circular table without intersecting, $\ldots$.

Contributed by Solomon W. Golomb
University of Southern California

Answer to Picture Puzzle

(Page 561)

Ed Begle

The Relationship between AB and BA

Charles R. Johnson and Erik A. Schreiner

Even when A and B are both n-by-n matrices, A and B need not commute, as we impress upon beginning students when matrix multiplication is first introduced. Moreover, as A and B need only be m-by-n and n-by-m, respectively, for both AB and BA to make sense, AB and BA need not even be the same size. Nonetheless, AB and BA are not independent of one another and, in fact, have much in common. For example, it is a simple calculation, requiring only matrix multiplication and manipulation of sums, that trace(AB) = trace(BA). From this simple fact, it follows immediately that

$$\operatorname{trace}\left([AB]^k\right) = \operatorname{trace}\left([BA]^k\right)$$

for any positive integer k. Thus, with $m \leq n$, the characteristic polynomial of BA is x^{n-m} times that of AB (see [**HJ**, problem 12, page 44]). It is not hard to see that even more is true. As

$$\begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix},$$

we see that

$$\begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \text{ is similar to } \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix},$$

from which it follows that the Jordan structure associated with the **nonzero** eigenvalues of AB is the same as that of BA. A simple example

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

however, shows that the eigenvalue 0 may have different Jordan structure in

$$AB = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ from that in } BA = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Of course, when $n > m$, BA has $n - m$ more zero eigenvalues than BA.

What then is the precise relationship between AB and BA? Actually, the somewhat subtly surprising answer was given some time ago in [**F**]. Let C and D be m-by-m and n-by-n complex matrices, respectively, with $m \leq n$. Then C may be written AB while D is BA if and only if (1) the Jordan structure associated with nonzero eigenvalues is identical in C and D and (2) if $m_1 \geq m_2 \geq \cdots$ are the sizes of Jordan blocks associated with 0 in C while $n_1 \geq n_2 \geq \cdots$ are the corresponding

sizes in D, then $|n_i - m_i| \le 1$ for all i. Otherwise, C and D are entirely independent. Here, for convenience, we fill out the lists of zero Jordan block sizes with 0's as needed.

Flander's proof of the surprising relationship between AB and BA deals with elementary divisors over a general field, is relatively abstract, and does not suggest the process of discovery, as was the style of the time. In the intervening 40 years, there has been little revisiting of the subject (see **[T]**, which focuses upon possibilities for rank and **[PM]** for variations), and the fundamental fact is not well known beyond experts. It is possible in today's explicitly constructive style to give a rather transparent proof, which requires only relatively basic background, and our proof may be modified to deal with elementary divisors over a general field. The stated relationship between AB and BA may be reduced to the following observation about a special Jordan form. Recall that a **nilpotent** matrix is a square one, each of whose eigenvalues is zero. Any other background needed may be found in **[HJ]**.

Lemma. *Let A be an n-by-n nilpotent matrix whose Jordan form consists of basic Jordan blocks of sizes $n_1 \ge n_2 \ge \cdots \ge n_p \ge 1$, $n_1 + \cdots + n_p = n$. If the Jordan form of the m-by-m nilpotent matrix*

$$B = \begin{bmatrix} A & X \\ 0 & 0 \end{bmatrix}$$

consists of basic Jordan blocks of sizes $m_1 \ge m_2 \ge \cdots \ge m_q \ge 1$, $m_1 + \cdots + m_q = m$, then $q \ge p$, $m_i = 1$ for $p+1 \le i \le q$, and $0 \le m_i - n_i \le 1$ for $i = 1, \dots, p$. Furthermore, any Jordan form meeting all these requirements may occur for some matrix of the form B.

We delay a proof of the lemma, which requires a bit of effort, until we see that Flanders' theorem is easily proved with it.

First, notice that $C = AB$ and $D = BA$ may be replaced by arbitrary and independent similarities, or, more importantly for the sake of argument, that A may be replaced by a matrix equivalent to it, i.e.,

$$C = AB \quad \text{and} \quad D = BA$$

if and only if

$$XCX^{-1} = (XAY)(Y^{-1}BX^{-1}) \quad \text{and} \quad Y^{-1}DY = (Y^{-1}BX^{-1})(XAY).$$

Thus, for necessity, it suffices to assume that

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

On the other hand, for sufficiency we may suppose that each of C and D is in any form achievable by similarity that we like.

To verify necessity, consider

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

partitioned conformably. Thus,

$$AB = \begin{bmatrix} B_{11} & B_{12} \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad BA = \begin{bmatrix} B_{11} & 0 \\ B_{21} & 0 \end{bmatrix}. \tag{1}$$

Since the Jordan structure associated with the nonzero eigenvalues of each comes entirely from B_{11}, it suffices to consider only the zero eigenvalues of B_{11} and to

assume that B_{11} is nilpotent. The lemma then applies to AB and also to BA, as transposition does not change Jordan structure. But, in each case the lemma says that the Jordan block sizes either increase by 1 or stay the same relative to the block sizes in B_{11}. In any event, the block sizes in AB and BA may then differ by at most one 1, as asserted.

To demonstrate sufficiency, suppose that C and D, meeting the stated requirements, are given. As the common nonzero parts of the Jordan form are easily attained by commuting direct summands, we may again assume that C and D are nilpotent. Now take A and B to be in the joint form used in the necessity proof. It is clear that, using the final assertion of the lemma, B_{11}, B_{12}, and B_{21} may be chosen to achieve C and D with the desired Jordan forms.

We now turn to proving the lemma. If $S^{-1}AS$ is in Jordan form, then similarity of B via $S \oplus I$ shows that we may take B to have the form

$$B = \left[\begin{array}{ccc|c} J_{n_1} & & 0 & X_1 \\ & \ddots & & \vdots \\ 0 & & J_{n_p} & X_p \\ \hline & 0 & & 0 \end{array}\right],$$

as $S^{-1}X$ runs through all possibilities if X does. Here J_t denotes the t-by-t nilpotent basic Jordan block, $n_1 + \cdots + n_p = n$, and

$$\begin{bmatrix} J_{n_1} & & 0 \\ & \ddots & \\ 0 & & J_{n_p} \end{bmatrix}$$

is the Jordan form of A. The lower right 0 block in B is k-by-k, with $k = m - n$, and the blocks X_i (which are n_i-by-k) form a partition of X. The lemma is then constructively verified in two stages. We first see that B is similar to a matrix of the same form in which each block X_i is a 0-1 matrix, with at most one 1, which if present, occurs in the last row, and any such 1's occur in different columns of X. The i-th Jordan block size in the Jordan form of B is then $n_i + 1$ if a 1 appears in X_i and n_i otherwise, $i = 1, \ldots, p$. Any additional Jordan blocks in B are 1-by-1. It is then clear that any set of Jordan blocks from A may be the ones to increase in size by 1 via choice of X.

In preparation for realizing the special form of B under similarity note that for $W = J_t^T X$,

$$\begin{bmatrix} I_t & W \\ 0 & I_k \end{bmatrix}\begin{bmatrix} J_t & X \\ 0 & 0 \end{bmatrix}\begin{bmatrix} I_t & -W \\ 0 & I_k \end{bmatrix} = \begin{bmatrix} J_t & Y \\ 0 & 0 \end{bmatrix}$$

in which $Y = X - J_t W = X - J_t J_t^T X$ has its first $t - 1$ rows equal to 0 and its last row the same as X, because $J_t J_t^T = I_{t-1} \oplus 0$. Performing a similarity upon B via

$$\left[\begin{array}{cccc|c} I_{n_1} & & & 0 & -W_1 \\ & I_{n_2} & & & -W_2 \\ & & \ddots & & \vdots \\ 0 & & & I_{n_p} & -W_p \\ \hline & & 0 & & -I_k \end{array}\right],$$

in which $W_i = J_{n_i}^T X_i$, we see that we may assume that in B the first $n_i - 1$ rows of each X_i are 0. The p-by-k matrix R formed by retaining the last row of each X_i is then critical. Imagine row reducing R, **without interchange of rows**, and then column reducing, so that the result R' is a 0–1 matrix with at most one 1 in each row and column. Each successive 1 may also be arranged to occur one column to the right of the prior one. The matrix B is then similar to one in which the rows R are replaced by the rows R'. The necessary row operations may be achieved by a sequence of similarities of the form

$$\begin{bmatrix} I & & 0 \\ & \ddots & \\ -U_{ij} & & \\ & & I \end{bmatrix}$$

in which $-U_{ij} = -[0 \;\; c_{ij}I]$ is in the i, j block position in a partition conformable with that of the Jordan form of A and is the only off-diagonal nonzero block. Here, the form of U_{ij} is necessary to preserve the Jordan form of A; it is important that the blocks J_{n_i} are arranged in decreasing order of size, and c_{ij} is the value necessary for a particular row operation. The sequence of similarities is determined in the following manner: Use the first nonzero entry of the top nonzero row of R as a pivot and zero out all entries of R in that column. Now move to the next nonzero row of R and use the first nonzero entry of this row as the next pivot. Observe that the new pivot may be in a column to the left of the previous pivot. The necessary column operations, achieved by right multiplication by an invertible k-by-k matrix V, may be implemented via similarity by $I_n \oplus V$.

We now have B in the asserted form. To see what the Jordan blocks for B are, we make the following more general observation. Let M be of the form

$$M = \left[\begin{array}{ccc|ccc} J_{n_1} & & 0 & X_{11} & \cdots & X_{1q} \\ & \ddots & & \vdots & \cdots & \vdots \\ 0 & & J_{n_p} & X_{p1} & \cdots & X_{pq} \\ \hline & & & J_{m_1} & & 0 \\ & 0 & & & \ddots & \\ & & & 0 & & J_{m_q} \end{array}\right].$$

If X_{ij} is such that $X_{ir} = 0$, $1 \le r \le q$, $r \ne j$ and $X_{sj} = 0$, $1 \le s \le p$, $s \ne i$, then M is permutation similar to

$$\begin{bmatrix} J_{n_i} & X_{ij} \\ 0 & J_{m_j} \end{bmatrix} \oplus M_1$$

in which M_1 is the submatrix of M resulting from removal of block row i from the upper half, block column i from the left half, block column j from the right half, and block row j from the lower half. Let X_{ij} be the j-th column of our transformed X_i. Since

$$\begin{bmatrix} J_{n_i} & X_{ij} \\ 0 & J_1 \end{bmatrix}$$

is J_{n_i+1} if $X_{ij} = [0, 0, \ldots, 0, 1]^T$ and is $J_{n_i} \oplus J_1$ if $X_{ij} = 0$, application of the preceding observation to the achieved special form of B proves the lemma.

A more detailed description of the technology available to determine the Jordan structure of a nilpotent block triangular matrix may be found in [**JS**].

REFERENCES

[F] H. Flanders, Elementary divisors of AB and BA, *Proc. Amer. Math. Soc.* 2 (1951), 871–874.

[HJ] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[JS] C. R. Johnson and E. A. Schreiner, Explicit Jordan form for certain block triangular matrices, II, *Linear Algebra Appl.* 164 (1992), 601–613.

[PM] W. V. Parker and B. E. Mitchell, Elementary Divisors of Certain Matrices, *Duke Math. J.* 19 (1952), 483–485.

[T] R. C. Thompson, On the Matrices AB and BA, *Linear Algebra Appl.* 1 (1968), 43–58.

Author's Note. This manuscript was essentially completed in 1991, shortly before the untimely death on Sept. 6, 1991 of the second author. This note is respectfully dedicated to the fond memory of him held by all his colleagues and collaborators.

Charles R. Johnson
Department of Mathematics
College of William and Mary
Williamsburg, VA 23185
crjohnso@cs.wm.edu

Erik A. Schreiner
Department of Mathematics and Statistics
Western Michigan University
Kalamazoo, MI 49008

On Thin Sets of Circles

Hajrudin Fejzić

Mathematicians are always amused by sets that are "small" in one sense and "big" in another. An example would be a set of plane measure 0 that contains a circle of every radius. Such a set was first given independently by Besicovitch and Rado and by Kinney. Kinney's set is very simple, but his proof that the set is of plane measure 0 is very complicated. Four years later Davies gave another set with the same properties, but with a relatively simple proof showing that his set is of plane measure 0. On the other hand, Davies' set is not so simple. Davies' proof relies on the fact that the so-called irregular 1-sets have projections of Lebesgue measure 0 in almost every direction. For the purpose of this note we briefly introduce the reader to these sets. More detailed definitions and further properties of these sets can be found in [**3**]. We say that a plane set is a 1-set if it has finite and positive linear measure. It is a remarkable fact that a 1-set either has projections of positive Lebesgue measure in all but possibly one direction or it has projections of Lebesgue measure 0 in almost every direction. Therefore if a 1-set has projections of Lebesgue measure 0 in two different directions it is irregular. Here we use the projection properties of irregular 1-sets to show that Kinney's set is of plane measure 0.

First we describe Kinney's set: Let C denote the Cantor set. It is well-known that $[0,1] \subset \{d - c | (d,c) \in C \times C\}$. Every $r \in (0,1)$ can be written in at least three different ways as a difference of two numbers in C. For each $r \in [0,1]$ we find the leftmost pair of points in $C \times C$ for which $r = d - c$. Let E denote the

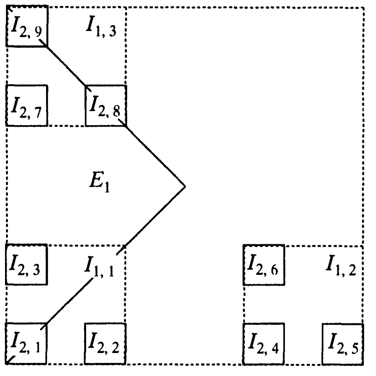


Figure 1

collection of such pairs. For $(c, d) \in E$, we construct a circle centered at $((c + d)/2, 0)$ and of radius $(d - c)/2$. It is clear that the union of all these circles contains circles of any radii between 0 and $1/2$. Kinney's set is this union. By taking countably many similar copies of Kinney's set, we get a set of plane measure 0 that contains a circle of every radius.

Theorem 1. *Kinney's set is of plane measure* 0.

Proof: First we show that there is a compact irregular 1-set, E_1, contained in $C \times C$, such that $E \subset E_1$. We construct E_1 in an infinite sequence of stages in the following way: In the first stage, we construct three squares, $I_{1,1}$, $I_{1,2} I_{1,3}$, of length $1/3$ in the two lower corners and in the upper-left corner of the unit square. In the second stage, construct 3^2 squares $I_{2,1}, \dots, I_{2,3^2}$, of length $1/3^2$ located in the two lower and the upper-left corner of each square from stage 1. Continue in this fashion, to get $\{I_{n,j}\}_{j=1,\dots,3^n}^{n=1,\dots,\infty}$. Let $G = \bigcap_{n=1}^{\infty}(\bigcup_{j=1}^{3^n} I_{n,j})$. Let E_1 denote the intersection of G and the triangle with vertices $(0,0)$, $(0,1)$, and $(\frac{1}{2}, \frac{1}{2})$. See Figure 1.

The set E_1 is obviously compact. Since the projection of E_1 onto the line $y = -x$ is the segment of length $\sqrt{2}/2$, $E \subset E_1$, and its linear measure, $\mathscr{H}^1(E_1)$, is greater than or equal to $\sqrt{2}/2$. On the other hand, E_1 can be covered by $(3^n + 1)/2$ squares of side $1/3^n$ for every n, so that

$$\mathscr{H}^1(E_1) \le \lim_{n \to \infty} \frac{3^n + 1}{2} \frac{1}{3^n} \sqrt{2} = \frac{\sqrt{2}}{2}.$$

Therefore $\mathscr{H}^1(E_1) = \sqrt{2}/2$ and E_1 is a 1-set. Since the projections of E_1 onto the x and y axes are of linear measure 0, E_1 is irregular.

Let $F: R^2 \to R^2$ be defined by $F(x, y) = (x + y, -xy)$. Since F is a real-analytic mapping, the set $A = F(E_1)$ is a compact irregular 1-set.

Let $C_{(p,q),r} = \{(u, v): u^2 + v^2 = 2up + 2vq + r^2 - p^2 - q^2\}$ denote the circle centered at the point (p, q) and with radius r. With this notation, Kinney's set, K, is

$$K = \bigcup_{(c,d) \in E} C_{((c+d)/2, 0), (d-c)/2} \subset \bigcup_{(c,d) \in E_1} C_{((c+d)/2, 0), (d-c)/2}.$$

For $(a, b) \in A$, let $D_{(a,b)} = \{(u, v): u^2 + v^2 = au + b\}$ and $D = \bigcup_A D_{(a,b)}$. It is easy to check that $K \subset D$.

Since A is compact, the set D is closed and is therefore measurable. We show that the plane measure of D is 0. To do so by Fubini's theorem it is enough to show that for almost every u the set of v's such that $(u, v) \in D$ is of linear measure 0. Since A is an irregular 1-set for almost every u, the set of projections of A onto a line with slope $-1/u$ is of linear measure 0. Hence for almost every u the set of y-intercepts of lines through points in A and with the fixed slope u is of linear measure 0. But $(u, v) \in D$ if and only if $u^2 + v^2$ is the y-intercept of a line, $k = ux + y$, with slope $-u$ that passes through some point $(a, b) \in A$. Therefore, for almost every u the set $\{(u^2 + v^2) : (u, v) \in D\}$ is of linear measure 0. It follows that for almost every u the set of v's such that $(u, v) \in D$ is of linear measure 0. See Figure 2.

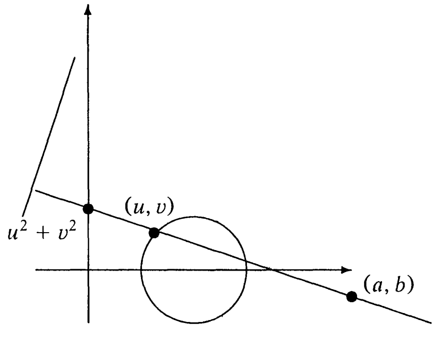


Figure 2

By Fubini's theorem, D is of plane measure 0, and since $K \subset D$, by the regularity of Lebesgue measure, K is measurable and K is of plane measure 0. □

Remark. By a similar argument, one can prove that if E is a compact irregular 1-set, then the set

$$K = \bigcup_{(p, r) \in E} C_{(p, 0), r}$$

is of plane measure 0. Now if the set E has the property that the projections of E onto the x-axis and onto the y-axis are the intervals $[0, 1] \times \{0\}$ and $\{0\} \times [0, 1]$, respectively, then K is a set of plane measure 0, which contains a circle of every radius between 0 and 1, and with centers of these circles filling the unit interval. By taking countably many similar copies of K we get a set of plane measure 0 that contains a circle of every radius and with centers of these circles filling the entire x-axis.

Question. Does there exist a set of plane measure 0 containing circles of every radius with centers on a parabola or any other "nice" curve other than a line? It seems that the techniques used in the proof of the theorem are not enough to answer this question.

REFERENCES

1. Besicovitch, A. S. & Rado, R. (1968). A plane set of measure zero containing circumferences of every radius, *J. London Math. Soc.* 43, 717–9.
2. Davies, R. O. (1972). Another thin set of circles, *J. London Math. Soc.* (2) 5, 191–2.

3. Falconer, K. J. (1985). *The geometry of fractal sets*. Cambridge University Press.
4. Kinney, J. R. (1968). A thin set of circles. *Amer. Math. Monthly* 75, 1077–81.

Department of Mathematics,
California State University San Bernardino,
San Bernardino, CA 92407
hfejzic@wiley.csusb.edu

A Proof of the Arithmetic Mean–Geometric Mean Inequality

Horst Alzer

The classical arithmetic mean–geometric mean inequality states that

$$G_n \equiv \prod_{i=1}^{n} a_i^{p_i} \leq \sum_{i=1}^{n} p_i a_i \equiv A_n \tag{1}$$

for all positive real numbers $a_1, \dots, a_n$ and $p_1, \dots, p_n$ with $p_1 + \cdots + p_n = 1$. Inequality (1), which is "probably the most important inequality, and certainly a keystone of the theory of inequalities" [**1**, p. 3] has found much interest among many mathematicians, and there are numerous new proofs, extensions, refinements, and variants of (1). We refer to the monographs [**1–5**] and the references therein. (In [**2**] more than 50 different proofs of (1) are given.) The purpose of this note is to present a short and simple proof of (1) that we could not locate in the literature.

We may assume that $a_1 \leq \cdots \leq a_n$. Then there exists an integer $k \in \{1, \dots, n-1\}$ such that $a_k \leq G_n \leq a_{k+1}$. This implies

$$\frac{A_n}{G_n} - 1 = \sum_{i=1}^{k} p_i \int_{a_i}^{G_n} \left(\frac{1}{t} - \frac{1}{G_n} \right) dt + \sum_{i=k+1}^{n} p_i \int_{G_n}^{a_i} \left(\frac{1}{G_n} - \frac{1}{t} \right) dt \geq 0, \tag{2}$$

since both sums have only non-negative terms. Moreover, equality holds in (2) if and only if $a_i = G_n$ for $i = 1, \dots, n$, that is, if and only if $a_1 = \cdots = a_n$.

REFERENCES

1. E. F. Beckenbach and R. Bellman, *Inequalities*, Springer, Berlin, 1983.
2. P. S. Bullen, D. S. Mitrinović, and P. M. Vasić, *Means and Their Inequalities*, Reidel, Dordrecht, 1988.
3. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, 1952.
4. D. S. Mitrinović, *Analytic Inequalities*, Springer, New York, 1970.
5. D. S. Mitrinović, J. E. Pečarić, and A. M. Fink, *Classical and New Inequalities in Analysis*, Kluwer, Dordrecht, 1993.

Morsbacher Strasse 10
51545 Waldbröl
Germany

Edited by: **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Development of Rigor in Mathematical Probability (1900–1950)

Joseph L. Doob

1 Introduction. This paper is a brief informal outline of the history of the introduction of rigour into mathematical probability in the first half of this century. Specific results are mentioned only in so far as they are important in the history of the logical development of mathematical probability.

The development of science is not a simple progression from one advance to the next. Judged by hindsight, the development is slow, proceeds in a zigzag course, with many wrong turns and blind alleys, and frequently moves in directions condemned by leading scientists. In the 1930's Banach spaces were sneered at as absurdly abstract, later it was the turn of locally convex spaces, and now it is the turn of nonstandard analysis. Mathematicians are no more eager than other humans to embrace new ideas, and full acceptance of mathematical probability was not realized until the second half of the century. In particular, many statisticians and probabilists resented the mathematization of probability by measure theory, and some still place mathematical probability outside analysis. The following quotations (in translation) are relevant.

Planck: *A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up with it.*

Poincaré: *Formerly, when one invented a new function, it was to further some practical purpose; today one invents them in order to make incorrect the reasoning of our fathers, and nothing more will ever be accomplished by these inventions.*

Hermite: (in a letter to Stieltjes) *I recoil with dismay and horror at this lamentable plague of functions which do not have derivatives.*

Probability theory began, and remained for a long time, an idealization and analysis of certain real life phenomena outside mathematics, but gradually, in the first half of this century, mathematical probability became a normal part of mathematics. The mathematization of probability required new ideas, and in particular required a new approach to the idea of acceptability of a function. In

Reprinted with kind permission of Birkhäuser Verlag AG, from *Development of Mathematics 1900–1950*, edited by Jean-Paul Pier, Basel, 1994, ISBN 0-8176-2821-5

view of the above quotations it is not surprising that acceptance of this mathematization was slow and faced resistance. In fact even now some probabilists fear that mathematization has removed the intrinsic charm from their subject. And they are right in the sense that the charm of the old, vague probability-mathematics, based on nonmathematical definitions, has split into two quite different charms: those of real world probability and of mathematical precision. But it must be stressed that many of the most essential results of mathematical probability have been suggested by the nonmathematical context of real world probability, which has never even had a universally acceptable definition. In fact the relation between real world probability and mathematical probability has been simultaneously the bane of and inspiration for the development of mathematical probability.

2 What is the real world (nonmathematical) problem? What is usually called (real world) probability arises in many contexts. Besides the obvious contexts of gambling games, of insurance, of statistical physics, there are such simple contexts as the following. Suppose an individual rides his bicycle to work. The rider would be surprised if, when the bicycle is parked, the valve on the front tire appeared in the upper half of the tire circle 10 successive days, just as surprised as if 10 successive tosses of a coin all gave heads. However, it is clear that (tire context) if the ride is very short, or (coin context) if the coin starts close to the coin landing place and the initial rotational velocity of the coin is low, the surprise would decrease and the probability context would become suspect. The moral is that the specific context must be examined closely before any probabilistic statement is made. If philosophy is relevant, an arguable question, it must be augmented by an examination of the physical context.

3 The law of large numbers. In a repetitive scheme of independent trials, such as coin tossing, what strikes one at once is what has been christened the *law of large numbers*. In the simple context of coin tossing it states that in some sense the number of heads in n tosses divided by n has limit $1/2$ as the number of tosses increases. The key words here are *in some sense*. If the law of large numbers is a mathematical theorem, that is, if there is a mathematical model for coin tossing, in which the law of large numbers is formulated as a mathematical theorem, either the theorem is true in one of the various mathematical limit concepts or it is not. On the other hand, if the law of large numbers is to be stated in a real world nonmathematical context, it is not at all clear that the limit concept can be formulated in a reasonable way. The most obvious difficulty is that in the real world only finitely many experiments can be performed in finite time. Anyone who tries to explain to students what happens when a coin is tossed mumbles words like *in the long run*, *tends*, *seems to cluster near*, and so on, in a desperate attempt to give form to a cloudy concept. Yet the fact is that anyone tossing a coin observes that for a modest number of coin tosses the number of heads in n tosses divided by n seems to be getting closer to $1/2$ as n increases. The simplest solution, adopted by a prominent Bayesian statistician, is the vacuous one: never discuss what happens when a coin is tossed. A more common equally satisfactory solution is to leave fuzzy the question of whether the context under discussion is or is not mathematics. Perhaps the fact that the assertion is called a *law* is an example of this fuzziness. The following statements have been made about this law (my emphasis):

Laplace: (1814) *This <u>theorem</u>, implied by common sense, was difficult to prove by analysis.*

Ville: (1939) *One sees no reason for this proposition to be true; but as it is impossible to prove experimentally that it is false, one can at least safely state it.*

Bauer: (translated from the context of dice to that of coins) *It is an experimentally established fact that the quotient... exhibits a deviation from* $1/2$ *which approaches* 0 *for large* n.

These statements illustrate the enduring charm of discussions of real world probability. Mathematicians, unfortunately, have felt forced to think about the following question, or at least to write about it.

4 What is probability? Here are some attempts to answer this question and to discuss the teaching of the subject.

Poincaré: (1912) *One can scarcely give a satisfactory definition of probability.*

Mazurkiewicz: (1915) *The theory of probability is not an independent element of mathematical instruction; nevertheless it is very desirable that a mathematician knows its general principles. Its fundamental concepts are incompletely determined. They contain many unsolved difficulties.*

v. Mises: (1919) *In fact, one can scarcely characterize the present state other than that probability is not a mathematical discipline.* (He proceeded to make it into a mathematical discipline by basing mathematical probability on a sequence of observations («Beobachtungen») with properties that cannot be satisfied by a mathematically well defined sequence. In a lighter mood he is said to have defined probability as a number between 0 and 1 about which nothing else is known.)

E. Pearson: (1935) (Oral communication) *Probability is so linked with statistics that, although it is possible to teach the two separately, such a project would be just a tour de force.*

Uspensky: (1937) In a useful textbook, he gave the following once common textbook definition. *If, consistent with condition* S, *there are* n *mutually exclusive, and equally likely cases, and* m *are favorable to the event* A, *then the mathematical probability of* A *is defined as* m/n.

The foregoing should make obvious the advisability of separating mathematical probability theory from its real world applications. Note however that no one doubts the real world applicability of mathematical probability. Gambling, genetics, insurance and statistical physics are here to stay.

Only *mathematical* probability will be discussed below, except for the following remark on coin tossing. Newtonian mechanics provides a partial mathematical model for coin tossing. In coin tossing, a solid body falls under the influence of gravity. Its motion is determined in Newton's model by his laws, and any discussion of what the coin does cannot be complete unless these laws are applied. Only these laws, rather than philosophical remarks, can explain the quantitative influence and importance of the initial and final conditions of the coin motion in order to justify allusions to equal likelihood of heads and tails. Of course these laws can at best reduce the analysis to considerations of the initial and final conditions of the toss, but these conditions can show what the «equal likelihoods» depends on and thereby give it a plausible interpretation.

5 Mathematical probability before the era of precise definitions. There were many important advances in mathematical probability before 1900, but the subject was not yet mathematics. Although nonmathematical probabilistic contexts suggested problems in combinatorics, difference equations and differential equations, there was a minimum of attention paid to the mathematical basis of the contexts, a maximum of attention to the pure mathematics problems they suggested. This unequal treatment was inevitable, because measure theory, needed for mathematical modeling of real world probabilistic contexts, had not yet been invented.

It was always clear that, however classical mathematical probability was to be developed, the concept of additivity of probability as applied to incompatible real world events was fundamental. Additive functions of sets were of course familiar to mathematicians from concepts of volume, mass and so on, long before 1900. It was realized that contexts involving averages led to probability. It was frequently clear how to use the contexts to suggest problems in analysis, but it was not clear how to formulate an overall mathematical context, that is, how to define a mathematical structure in which the various contexts could be placed.

A weaker condition than additivity was less familiar but turned out to be essential later. The standard loose language will be used here. If $x_1, x_2, \ldots$ are numbers obtained by chance, and if A is a set of numbers, consider the probability that at least one of the members of this sequence lies in A, or, in more colorful language, consider the probability that an orbit of this motion through points of a line hits A. The usual calculation (ignoring here all notions of rigour), defines a function $A \to \phi(A)$ which in general is not additive. In fact ϕ satisfies the inequality

$$\phi(A) + \phi(B) - \phi(A \cup B) \geq \phi(A \cap B), \tag{5.1}$$

whereas additivity of ϕ would imply equality in (5.1). The point is that the left side of (5.1) is the probability that the sequence $x\bullet$ hits both A and B, a probability at least equal to, and in general greater than, $\phi(A \cap B)$, the probability that the sequence hits $A \cap B$. The inequality (5.1), the *strong subadditivity* inequality, is satisfied also by the electrostatic capacity of a body in $\mathbf{R}^3$, and this fact hints at the close connection between potential theory and probability, developed in great detail in the second half of the century with the help of Choquet's theory of mathematical capacity.

6 The development of measure theory. Recall that a *Borel field* ($= \sigma$ algebra) of subsets of a space is a collection of subsets which is closed under the operations of complementation and the formation of countable unions and intersections. The class of *Borel sets* of a metric space is the smallest set σ algebra containing the open sets of the space. A *measurable space* is a pair, $(S, \mathbb{S})$, where S is a space and $\mathbb{S}$ is a σ algebra of subsets of S. The sets of $\mathbb{S}$ are the *measurable sets* of the space. In the following, if S is metric, the coupled σ algebra making it into a measurable space will always be the σ algebra of its Borel sets. In particular $(\mathbf{R}^N, \mathbb{R}^N)$ denotes N dimensional Euclidean space coupled with its Borel sets. The superscript will be omitted when $N = 1$. A measurable function from a measurable space $(S_1, \mathbb{S}_1)$ into a measurable space $(S_2, \mathbb{S}_2)$ is a function from S_1 into S_2 with the property that the inverse image of a set in $\mathbb{S}_2$ is a set in $\mathbb{S}_1$.

Measure theory started with Lebesgue's thesis (1902), which extended the definition of volume in $\mathbf{R}^N$ to the Borel sets. Radon (1913) made the further step

to general measures of Borel sets of $\mathbf{R}^N$ (finite on compact sets). These measures are usually extended to slightly larger classes than the class of Borel sets, by *completion*. Finally Fréchet (1915), 13 years after Lebesgue's thesis, pointed out that all that the usual definitions and operations of measure theory require is a σ algebra of subsets of an abstract space on which a measure, that is, a positive countably additive set function, is defined. At each step of this progression not necessarily positive countably additive set functions—signed measures—were incorporated into the theory. As noted below, the Radon-Nikodym theorem (1930), which gives conditions necessary and sufficient that a countably additive function of sets can be expressed as an integral over the sets, turned out to be the final essential result needed to formulate the basic mathematical probability definitions. Thus it was 28 years before Lebesgue's theory was extended far enough to be adequate for the mathematical basis of probability. This extension was not developed in order to provide a basis for probability, however. Measure theory was developed as a part of classical analysis, and applications in analysis were immediate, for example to the (Lebesgue measure) almost everywhere derivability of a monotone function.

There has been criticism of the fact that mathematical probability is usually prescribed not only to be additive but even to be countably additive. The question whether real world probability is countably additive, if the question is to be meaningful, asks whether a mathematical model of real world probabilistic phenomena *necessarily* always involves *countably* additive set functions. In fact there may well be real world contexts for which the appropriate mathematical model is based on finitely but not countably additive set functions. But there have been very few applications of such set functions in either mathematical or nonmathematical contexts, and such set functions will not be discussed further here.

7 Early applications of explicit measure theory to probability. Some probabilistic slang will be needed, enduring relics of the historical background of probability theory. A probability space is a triple $(S, \mathbb{S}, \mathrm{P})$, where $(S, \mathbb{S})$ is a measurable space and P is a measure on $\mathbb{S}$ with $\mathrm{P}(S) = 1$. A measure with this normalization is a *probability measure*. A *random variable* is a measurable function from a probability space $(S, \mathbb{S}, \mathrm{P})$, into a measurable space $(S', \mathbb{S}')$. The space S', or, when one writes carefully, $(S', \mathbb{S}')$, is the *state space* of the random variable. Mutual independence of random variables is defined in the classical way. The *distribution* of a random variable x is the measure P_x on $\mathbb{S}'$ defined by setting

$$P_x(A') = P\{s \in S : x(s) \in A'\}.$$

The joint distribution of finitely many random variables defined on the same probability space is obtained by making x into a vector and specifying $\mathbb{S}'$ and S' correspondingly. A stochastic process is a family of random variables $\{x(t,\bullet), t \in \mathcal{I}\}$ from some probability space $(S, \mathbb{S}, \mathrm{P})$, into a state space $(S', \mathbb{S}')$. The set $\mathcal{I}$ is the *index set* of the process. Thus a stochastic process defines a function of two variables, $(t, s) \to x(t, s)$, from $\mathcal{I} \times S$ into the state space. The function $x(t,\bullet)$ from S into S' is the tth random variable of the process; the function $x(\bullet, s)$ from $\mathcal{I}$ into S' is the sth sample function, or sample path, or sample sequence if $\mathcal{I}$ is a sequence.

Borel (1909) pointed out that in the dyadic representation $x = x_1x_2\ldots$ of a number x between 0 and 1, in which each digit x_j is either 0 or 1, these digits are functions of x, and if the interval $[0, 1]$ is provided with Lebesgue measure, a probability measure on this interval, these functions miraculously become random

variables which have exactly the distributions used in calculating coin tossing probabilities. That is, 2^{-n} is the probability assigned to the event that, in a tossing experiment, the first n tosses yield a specified sequence of heads and tails, and 2^{-n} is also the total length (= Lebesgue measure) of the finite set of intervals whose points x have dyadic representations with a specified sequence of 0's and 1's in the first n places. Thus a mathematical version of the law of large numbers in the coin tossing context is the existence in some sense of a limit of the sequence of function averages $\{(x_1 + \cdots + x_n)/n,\ n \geq 1\}$. Classical elementary probability calculations imply that this sequence of averages converges in measure to $1/2$, but a stronger mathematical version of the law of large numbers was the fact deduced by Borel—in an unmendably faulty proof—that this sequence of averages converges to $1/2$ for (Lebesgue measure) almost every value of x. A correct proof was given a year later by Faber, and much simpler proofs have been given since. [Fréchet remarked tactfully: «Borel's proof is excessively short. It omits several intermediate arguments and assumes certain results without proof.»] This theorem was an important step, an example of a new kind of convergence theorem in probability. Observe that (fortunately) pure mathematicians need not interpret this theorem in the real world of real people tossing real coins. Some of the quotations given above indicate that they not only need not but should not.

Daniell (1918) used a deep approach to measure theory in which integrals are defined before measures to get a (rather clumsy) approach to infinite sequences of random variables by way of measures in infinite dimensional Euclidean space.

The Brownian motion stochastic process in $\mathbf{R}^3$ is the mathematical model of Brownian motion, the motion of a microscopic particle in a fluid as the particle is hit by the molecules of the fluid. The process is normalized by supposing it starts at the origin of a cartesian coordinate system in $\mathbf{R}^3$, and a (normalized) Brownian motion process in $\mathbf{R}$ is the process of a coordinate function of a normalized process in $\mathbf{R}^3$, vanishing initially. A (normalized) Brownian motion process in $\mathbf{R}^N$ is a process defined by N mutually independent Brownian motion processes in $\mathbf{R}$. It was well known what the joint distributions of the random variables of a Brownian motion process should be, and it had been taken for granted that in a proper mathematical model the class of continuous paths would have probability 1. By 1900, Bachelier had even derived various important distributions related to the Brownian motion process in $\mathbf{R}$, such as that of the maximum change during a time interval, by finding corresponding distributions for a certain discrete random walk and then going to the limit as the walk steps tended to 0. More precisely, what Bachelier derived were distributions valid for a Brownian motion process if in fact there was such a thing as a Brownian motion process, and if it was approximable by his random walks. Observe that there was no question about the existence of Brownian motion; Brownian motion is observable under a microscope. But there was as yet no proof of the existence of a stochastic process, a mathematical construct, with the desired properties. Wiener (1923) constructed the desired Brownian motion process, now sometimes called the *Wiener process*, by applying the Daniell approach to measure theory to obtain a measure with the desired properties on a space S of continuous functions: if $x(t,\bullet)$ is the random variable defined by the value at time t of a function in S, the stochastic process of these random variables is a stochastic process with sample functions the members of S, and with the joint random variable distributions those prescribed for the Brownian motion process.

Bachelier's results remained unnoticed for years, and in fact were rediscovered several times. Wiener's work, like his fundamental work in potential theory, had

little immediate influence because it was published in a journal which was not widely distributed. It was an aspect of his genius that he carried out his Brownian motion research then and later without knowledge of the slang and some of the useful elementary mathematical techniques of probability theory.

Steinhaus (1930) demonstrated that classical arguments to derive standard probability theorems could be placed in a rigorous context by taking Lebesgue measure on a linear interval of length 1 as the basic probability measure, interpreting random variables as Lebesgue measurable functions on this interval, and expectations of random variables as their integrals. No new proofs were required; all that was required was a proper translation of the classical terminology into his context. If this were all mathematization of probability by measure theory had to offer, the scorn of rigorous mathematics expressed by some nonmathematicians would be justified.

8 Kolmogorov's 1933 monograph. Kolmogorov (1933) constructed the following mathematical basis for probability theory.

(a) The context of mathematical probability is a probability space $(S, \mathbb{S}, \mathrm{P})$. The sets in $\mathbb{S}$ are the mathematical counterparts of real world events; the points of S are counterparts of elementary events, that is of individual (possible) real world observations.

(b) Random variables on $(S, \mathbb{S}, \mathrm{P})$, are the counterparts of functions of real world observations. Suppose $\{x(t,\bullet), t \in \mathscr{I}\}$ is a stochastic process on a probability space $(S, \mathbb{S}, \mathrm{P})$, with state space S'. A set of n of the process random variables has a probability distribution on S'^n. Such finite dimensional distributions are mutually compatible in the sense that if $1 \leq m < n$, the joint distribution of $x(t_1,\bullet), \ldots, x(t_m,\bullet)$ on S'^m is the m-dimensional distribution induced by the n-dimensional distribution of $x(t_1,\bullet), \ldots,$ $x(t_n,\bullet)$ on S'^n.

(c) Conversely, Kolmogorov proved that given an arbitrary index set $\mathscr{I}$, and a suitably restricted measurable space $(S'', \mathbb{S}')$ (for example, the measurable space can be a complete separable metric space together with the σ algebra of its Borel sets) and a mutually compatible set of distributions on S'^n, for integers $n \geq 1$, indexed by the finite subsets of $\mathscr{I}$, there is a probability space and a stochastic process $\{x(t,\bullet), t \in \mathscr{I}\}$ defined on it, with state space S', with the assigned joint random variable distributions. To prove this result he constructed a probability measure on a σ algebra of subsets of the product space $S'^{\mathscr{I}}$, the space of all functions from $\mathscr{I}$ into S', and obtained the required random variables as the coordinate functions of $S'^{\mathscr{I}}$.

(e) The expectation of a numerically valued integrable random variable is its integral with respect to the given probability measure.

(f) The classical definition of the conditional probability of an event (measurable set) A, given an event B of strictly positive probability, is $\mathrm{P}(A \cap B)/P(B)$. In this way, for fixed B, new probabilities are obtained, and expectations of random variables for given B are computed in terms of these new *conditional* probabilities. More generally, given an arbitrary collection of random variables, conditional probabilities and expectations relative to given values of those random variables are needed, functions of the values assigned to the conditioning random variables. If $(S, \mathbb{S}, \mathrm{P})$ is a probability space, and if a collection of random variables is given, let $\mathbb{F}$ be the smallest sub σ algebra of $\mathbb{S}$ relative to which all the given random variables are

measurable. This σ algebra is the *σ algebra generated by conditions imposed on the given random variables*. A reasonable interpretation of a measurable real valued function of the given collection of random variables is a measurable function from $(S, \mathbb{F})$ into $\mathbf{R}$. The Kolmogorov *conditional expectation* of a real valued integrable random variable x on $(S, \mathbb{S}, \mathrm{P})$, relative to a σ algebra $\mathbb{C}$ of measurable sets, is a random variable which is measurable relative to $\mathbb{C}$ and has the same integral as x on every set in $\mathbb{C}$. The existence of such a random variable, and its uniqueness up to P-null sets, is assured by the Radon-Nikodym theorem. The conditional expectation of x relative to a collection of random variables is defined as the conditional expectation of x relative to the σ algebra generated by conditions on the random variables. A conditional probability of a measurable set A is defined as the conditional expectation of the random variable which is 1 on A and 0 elsewhere.

Kolmogorov's 1933 exposition paints a discouraging picture of mathematical progress. In the first pages of his monograph he states explicitly that real valued random variables are measurable functions and expectations are their integrals. Even as late as 1933, however, he must have thought that mathematicians were not familiar with measure theory. In fact in the body of his monograph, when he comes to the definition of a real valued random variable, he does not simply refer back to the first pages of the monograph and say that a random variable is a measurable function. Instead he actually defines measurability of a real valued function, and similarly when he defines the expected value of a random variable he does not simply state that it is the integral of the random variable with respect to the given probability measure, but he actually defines the integral. Later in the monograph, when he needs Lebesgue's theorem allowing taking limits of convergent function sequences under the sign of integration, he does not simply refer to Lebesgue but gives a detailed proof of what he needs. As confirmation of Kolmogorov's caution in invoking measure theory, the author recalls his student experience in 1932 when there were professorial disapproving remarks on the extreme generality of a seminar lecture given by Saks on what is now called the Vitali-Hahn-Saks theorem, a theorem which has since become an important tool in probability theory. [He also recalls that he did not understand the point of Kolmogorov's measure on a function space until long after he had read the monograph.]

It was some time before Kolmogorov's basis was accepted by probabilists. The idea that a (mathematical) random variable is simply a function, with no romantic connotation, seemed rather humiliating to some probabilists. A prominent statistician in 1935 wondered whether two orthogonal real valued random variables with zero means (integrals) are necessarily independent, as they are under the added hypothesis that they have a bivariate Gaussian distribution. He was rather surprised by the example of the sine and cosine functions on the interval $[0, 2\pi]$, with probability measure defined as Lebesgue measure divided by 2π. These two functions, orthogonal and with zero means but not independent, are not the kind of random variables probabilists were used to. Some analysts may be gratified, some humiliated, to learn that in discussing Fourier series they can be accused of discussing probabilities and expectations.

9 Expansion backwards of the Kolmogorov basis. Kolmogorov's basis for mathematical probability can be expanded, and should be expanded in the view of some probabilists, who want to start with some not necessarily numerical mathematical version of the confidence of observers that certain events will occur, and to

proceed postulationally to numerical evaluations of this confidence, and finally to additivity. Such an analysis may be enlightened in discussing the appropriateness of mathematical probability as a model for real world phenomena, but any approach to the subject which ends with a justification of the classical calculations and is mathematically usable, will end with Kolmogorov's basis, however phrased, because all the measure basis to probability does is to give a formal precise mathematical framework for the classical calculations and their present refinements. This framework had made it possible to apply mathematical probability in many other mathematical fields, for example to potential theory and partial differential equations. Although such applications were made in the past before the acceptance of measure theory as the basis of probability, the probabilistic context served only to suggest mathematics and was not an integral part of the mathematics. The meaning of solutions as probabilities and expectations could not be formulated and exploited.

10 Uncountable index sets. If the index set $\mathscr{I}$ of a stochastic process $\{x(t, \bullet), t \in \mathscr{I}\}$ is an interval of the line, and if the state space of the random variables is **R**, the class of continuous sample functions may not be measurable. This difficulty arises in the processes derived by the Kolmogorov construction of a measure on a function space, for example, whatever the choice of joint distributions of the process random variables. To understand the difficulty, observe that if the index set $\mathscr{I}$ of a stochastic process with state space **R** is an interval, and if $\mathscr{J}$ is a subset of $\mathscr{I}$, the function $s \to \sup_{t \in \mathscr{J}} x(t, s)$ is measurable if $\mathscr{J}$ is countable, but need not be measurable if $\mathscr{J}$ is uncountable. If boundedness and continuity of sample functions are to be discussed, some modification of the probability relations of the random variables of a stochastic process should be devised to make such suprema measurable functions. A clumsy approach was proposed by Doob (1937) but a more usable one was not devised until after 1950.

11 Reluctance to accept measure theory by probabilists. There was considerable resistance to the acceptance and exploitation of measure theory by probabilists, both in Kolmogorov's day and later. The following quotation is an example of the reluctance of some mathematicians to separate the mathematics from the context that inspired it.

Kac (1959) *How much fuss over measure theory is necessary for probability theory is a matter of taste. Personally I prefer as little fuss as possible because I firmly believe that probability theory is more closely related to analysis, physics and statistics than to measure theory as such.*

12 New relations between functions made possible by the mathematization of probability. Probability theory suggested new relations between functions. For example consider the sequence $x_1, x_2, \ldots$ of real valued integrable random variables on a probability space $(S, \mathbb{S}, \mathrm{P})$ and suppose that the conditional expectation of x_n given $x_1, \ldots, x_{n-1}$ vanishes (P) almost everywhere, for $n > 1$, that is, the integral of x_n over any set determined by conditions on the preceding random variables vanishes. If these random variables are square integrable, this condition is equivalent to the condition, much stronger than mutual orthogonality, that x_n is orthogonal to every square integrable function of $x_1, \ldots, x_n$. Bernstein (1927) seems to have been the first to treat such sequences systematically. This condition on a sequence of functions means that in a reasonable sense the sequence of partial sums of the given sequence is the counterpart of a fair game. In fact, the partial sums $y_1, y_2, \ldots$ are characterized by the property that the expectation of y_n

relative to $y_1, \ldots, y_{n-1}$ is equal to y_{n-1} almost everywhere on the probability space. Processes with this property, called *martingales*, first used explicitly by Ville (1939), have had many applications, for example to partial differential equations, to derivation, and to potential theory. Another important class of sequences of random variables is the class of sequences with the Markov property. These sequences are characterized by the fact that when $n \geq 1$ the conditional probabilities for x_n relative to $x_1, \ldots, x_{n-1}$ are equal almost everywhere to those for x_n relative to x_{n-1}. Roughly speaking, the influence of the present, given the past, depends only on the immediate past. The Markov property, introduced in a very special case by Markov in 1906, (named in his honor by others) has proved very fruitful, for example, leading in the second half of the century to a probabilistic potential theory, generalizing and including classical potential theory.

13 What is the place of probability theory in measure theory, and more generally in analysis? It is considered by some mathematicians that if one deals with analytic properties of probabilities and expectations then the subject is part of analysis, but that if one deals with sample sequences and sample functions then the subject is probability but not analysis. These authors are in the interesting situation that in considering a function of two variables, $(t, s) \to x(t, s)$—as in considering stochastic processes—they call it analysis if the family of functions $x(t,\bullet)$ as t varies is studied, but call it probability and definitely not analysis if the family of functions $x(\bullet, s)$ as s varies is studied. More precisely, they regard discussions of distributions and associated questions as analysis, but not discussions in terms of sample functions. This point of view is expressed in the following quotation.

Protter *By developing his integral in* 1944 *with stochastic processes as integrands, Itô was able to study multidimensional diffusions with purely probabilistic techniques, an improvement over the analytic methods of Feller.*

The following remark on the convergence of a sum of orthogonal functions illustrates the difficulty in separating (mathematical) probability from the rest of analysis. The measure space is a probability space, but with trivial changes the discussion is valid for any finite measure space.

If $x\bullet$ is an orthogonal sequence of functions, on a probability measure space, and if x_n^2 has integral σ_n^2, then (Riesz-Fischer) Σx_n converges in the mean if

$$\sum \sigma_n^2 < +\infty. \tag{13.1}$$

The orthogonal series converges almost everywhere if either (Menšov-Rademacher) (13.1) is strengthened to

$$\sum \sigma_n^2 \log^2 n < +\infty, \tag{13.2}$$

or (Lévy, 1937) the condition (13.1) is kept but the orthogonality condition is strengthened to the condition in Section 12.

The reader should judge which of these results is measure theoretic and which is probabilistic, whether there is any point in evicting mathematical probability from analysis, and if so whether measure theory should also be evicted.

101 West Windsor Road, Apt. 1104
Urbana, Illinois 61801-6663
doob@symcom.math.uiuc.edu

THE AUTHORS

PAULO RIBENBOIM—a friend of numbers—was born in Brazil (1928). He spent six years, with Jean Dieudonné in Nancy, and with Wolfgang Krull in Bonn and received the Ph.D. in Sao Paulo. Fulbright fellowship at the University of Illinois in Urbana and since 1962, professor at Queen's University in Kingston, Canada. Honorary degree by the Universite of Caen (France), Fellow of the Royal Society of Canada and author of several books, including *13 Lectures on Fermat's Last Theorem*, *The Book of Prime Number Records*, *L'Arithmetique des Corps*, *Catalan's Conjecture*.

ROGER PINKHAM was born in Maine of seafaring people, went to the University of Maine as an undergraduate, and got his Ph.D. from Harvard. Consulting jobs with Arthur D. Little, Inc. and Bell Telephone Laboratories broadened his tastes and insights, while a post-doc at Princeton under J. W. Tukey immeasurably sharpened his skills. Mathematics and statistics are life-long passions from which he hopes never to recover, and he is blessed with a wife and friends, second to none.

CHARLES GROETSCH received his Ph.D. from LSU in 1971. Except for visiting positions in the U.S., Europe, and Australia, he has been at the University of Cincinnati ever since. From 1985 to 1990 he served as Head of the Department of Mathematical Sciences at U.C. His research is focused on the theory of approximate methods for ill-posed inverse problems and he is the author of several books on themes related to this topic. Groetsch is a SIAM Visiting Lecturer and a recipient of the MAA's *George Polya Award.* When not teaching and trying to do mathematics, he spends more time than he should trying to get friends, colleagues, and total strangers help him make his computer work.

DONALD P. MINASSIAN received his bachelor's in social studies in 1957 from Fresno State University, his master's in mathematics in 1963 from Brown and in 1965 from Michigan, and his doctorate in mathematics from Michigan in 1967. (Thesis area: ordered groups). He also studied graduate anthropology at Berkeley. From 1967 he has taught at Butler University, where he is Professor, in mathematics and more recently in actuarial science, too. He became Associate of the Society of Actuaries in 1981, Member of the American Academy of Actuaries in 1983, and Fellow of the Society of Actuaries in 1993. He also passed the Uniform CPA Examination in 1976, received his J.D. from Indiana University, Indianapolis, in 1986, and passed the Indiana Bar in 1987. He belongs to Phi Beta Kappa and other honorary societies. He has worked as an actuary in both casualty and life insurance companies. His most recent interest is stochastic interest rates. On request he will mail jokes—including atrocious puns—about mathematicians, anthropologists, actuaries, accountants, lawyers, or college professors.

T. Y. LAM received his Ph.D. from Columbia University under the direction of H. Bass, and has taught at U.C. Berkeley since 1968. He was Sloan Fellow in 1972–74, Guggenheim Fellow in 1981–82, and received the Leroy P. Steele Prize for mathematical exposition from the American Mathematical Society in 1982. He has diverse interests in algebra, and has published in algebraic K-theory, group theory and representations, ring theory, field theory, quadratic forms, number theory, and combinatorics. Lam is an avid collector of mathematics books, and enjoys meeting hundreds of mathematicians in his current capacity as Deputy Director of the Mathematical Sciences Research Institute at Berkeley.

KA HIN LEUNG received his B.Sc. from the University of Hong Kong and his Ph.D. from the University of California, Berkeley, under the guidance of T. Y. Lam. After graduation, he joined the faculty of the National University of Singapore, where he is now a Senior Lecturer. His research interests are quadratic forms, combinatoric designs, ordered division rings, and $*$-fields. He also enjoys teaching mathematics to his two sons, Ngai Hang and Wing Hang.

EDWARD B. BURGER received his B.A. from Connecticut College in 1985 and his Ph.D. from the University of Texas at Austin in 1990 under the supervision of Jeffrey D. Vaaler. He was a Postdoctoral Fellow at the University of Waterloo before arriving at Williams College. Recently, he spent part of a sabbatical leave at the University of Colorado at Boulder. His research interests are in diophantine analysis and geometry of numbers. When he is not thinking about number theory, he is either giving entertaining math performances, making mathematical guest appearances on the NBC-TV News in Albany, NY, writing comedy plays, or studying the collected works of Jack Benny and Johnny Carson.

THOMAS STRUPPECK is a professional problem solver and a subprofessional bridge player. He did undergraduate work at Tulane and doctoral work at the University of Texas; nonetheless these institutions granted him degrees. He has taught at Rutgers and UT, and at various times has been a programmer, consultant, cartographer, and systems analyst. He is now an actuarial analyst at Republic Insurance, where many of these experiences have proved useful; however, no application of p-adic analysis has arisen yet.

LEONARD GILLMAN was a Juilliard fellow in piano for five years before turning to mathematics. After nine years in naval operations research, he completed a Columbia Ph.D. in transfinite numbers (age 36); taught at Purdue, Rochester, and Texas, with two years at the Institute for Advanced Study, one as a Guggenheim fellow (silver anniversary of Juilliard); retired in 1987. He is coauthor with Meyer Jerison of *Rings of Continuous Functions*. He was an AMS Associate Secretary for two years, MAA treasurer for 13, president for the canonical two, and pianist at four national meetings, two each with Louis Rowen, cello (1976 and 1980) and William Browder, flute (1989, Presidents' concert) and 1992 (Past-Presidents' concert).

FREEMAN DYSON was born in England in 1923, studied mathematics at Cambridge with Besicovitch, Hardy, and Littlewood, and published papers in number theory and topology. In 1947 he switched to physics and studied at Cornell with Bethe and Feynman. After thirty years as a physicist, he switched to writing books about science for the general public, following the example of Hardy, who once said to him, "Young men should prove theorems, old men should write books".

Finding Herndon reading a new book, *The Annual of Science*, [Abraham Lincoln] glanced through it and commented that the book was on the right track because it took account of failures as well as successes in its field. "Too often we read only of successful experiments in science and philosophy, whereas if the history of failure and defeat was included there would be a saving of brainwork as well as time. The evidence of defeat, the recital of what was not as well as what cannot be done serves to put the scientist or philosopher on his guard—sets him to thinking on the right line."

These remarks were prophetic, in their way, for Lincoln had arrived earlier than usual one morning at the office. Spread before him on his desk were sheets of paper covered with figures and equations, plenty of blank paper, a compass, rule, pencils, bottles of ink in different colors. He hardly turned his head as Herndon came in. He covered sheet after sheet with more figures, signs, symbols. As he left for the courthouse later in the day he told Herndon he was trying to square the circle.

He was gone only a short time, came back and spent the rest of the day trying to square the circle, and next day again toiled on the famous problem that has immemorially baffled mathematicians. After two days' struggle, worn down physically and mentally, he gave up trying to square the circle.

From *Abraham Lincoln*, Vol. 1, by Carl Sandburg
Harcourt, Brace, 1926, p. 476.

Contributed by W. A. Beyer
Los Alamos National Laboratory

10542. *Proposed by Jean Anglesio, Garches, France.*

Let $\mathcal{C}$ be the circumcircle of a triangle $A_0B_0C_0$ and $\mathcal{I}$ the incircle. It is known that, for each point A on $\mathcal{C}$, there is a triangle ABC having $\mathcal{C}$ for circumcircle and $\mathcal{I}$ for incircle. Show that the locus of the centroid G of triangle ABC is a circle that is traversed three times by G as A traverses $\mathcal{C}$ once, and determine the center and radius of this circle.

NOTES

(**10538**) Two versions of this problem, one by the first four named authors, and one by the last two, arrived within a short time. The similarity of the statements suggested that they be combined. The proposers noted that the case $n = 2$ has appeared on various national selection tests for the International Mathematical Olympiad. (**10542**) The existence of ABC is a special case of Poncelet's theorem. Details may be found in M. Berger, *Geometry I*, Springer-Verlag, 1987, p. 316

SOLUTIONS

The Superregular Graphs

6617 [1989, 942]. *Proposed by Andrew Vince, University of Florida, Gainesville, FL.*

A graph Γ is *regular* if each vertex has the same degree. For a vertex x let Γ_x and Δ_x denote the subgraphs of $\Gamma - x$ induced by the vertices adjacent to and nonadjacent to x, respectively. Define *superregular* recursively as follow. The empty graph is superregular and Γ is superregular if Γ is regular and both Γ_x and Δ_x are superregular for all x. Characterize the superregular graphs.

Solution by Randall B. Maddox, Pepperdine University, Malibu, CA. We adopt the following notation: K_n is the complete graph on n vertices, mK_n is m disjoint copies of K_n, C_n is the cycle on n vertices, and G_n is the graph whose vertex set consists of n^2 vertices arranged in an $n \times n$ square, with two vertices adjacent if and only if they are in the same row or column of the square. (The graph G_n is otherwise known as the Cartesian product of K_n with itself.) For any graph G, we let $\overline{G}$ denote its complement.

The superregular graphs are precisely the following: C_5, mK_n $(m, n \geq 1)$, G_n $(n \geq 1)$, and the complements of these graphs. Call this class of graphs S.

Theorem 1. *Every graph in S is superregular.*

Proof. It is easy to see that if a graph Γ is superregular, then so is its complement $\overline{\Gamma}$. The graph $\Gamma = K_n$ is certainly superregular, since for any x, $\Gamma_x = K_{n-1}$, which is superregular by induction, and Δ_x is empty. The graph $\Gamma = mK_n$ is then seen to be superregular, since for any x, $\Gamma_x = K_{n-1}$, which is superregular, and $\Delta_x = (m-1)K_n$, which is superregular by induction on m. The graph $\Gamma = G_n$ is superregular, since for any x, $\Gamma_x = 2K_{n-1}$, which is superregular, and $\Delta_x = G_{n-1}$, which is superregular by induction. Finally C_5 is easily seen to be superregular.

What remains is to prove the converse. Before addressing this, we first point out two basic properties of superregular graphs.

Proposition 2. *If* Γ *is a connected, superregular graph, then any pair of nonadjacent vertices of* Γ *have a common neighbor. If* Γ *is superregular and not connected, then* $\Gamma = mK_n$ *for some* m *and* n.

Proof. Suppose that Γ is a superregular graph of degree r and that w, x, y, z is a path in Γ such that the distance from w to z is 3. Then $y, z \in \Delta_w$, y has no more than $r-1$ neighbors in Δ_w, and z has exactly r neighbors in Δ_w. The same conclusion holds if w, x, y is a path in Γ such that the distance from w to y is 2, and z lies in a different component. Thus, in either case, Δ_w is not regular, which is a contradiction.

The remainder of the solution is devoted to the proof of the converse of Theorem 1.

Theorem 3. *Every superregular graph is in* S.

Proof. Suppose, in order to obtain a contradiction, that Γ is a superregular graph not in S and that, among all such graphs, Γ has the fewest vertices. Then Γ_x and Δ_x must be in S. The various possibilities for Γ_x will be considered in the following propositions. Note that, by Proposition 2, we may assume that Γ is connected.

Proposition 4. *For all* x, Γ_x *is not* C_5.

Proof. Suppose $\Gamma_x = C_5$. Then Γ is 5-regular, so every vertex in Γ_x has exactly 2 neighbors in Δ_x. Thus there are exactly 10 edges between Γ_x and Δ_x. It follows that Δ_x has a number of vertices which divides 10. Since this number must be even (the 5-regular graph Γ must have an even number of vertices), it must be 2 or 10. But if it is 2, then Γ must be $\overline{K_3}$ joined to C_5 with all possible edges, which is not superregular. So Δ_x must have 10 vertices and be 4-regular. The only such graph in S is $2K_5$, but then again Γ is not superregular.

Proposition 5. *For all* x, Γ_x *is not* $\overline{G_n}$.

Proof. We prove the stronger result that if Γ is any superregular graph with $\Gamma_x = \overline{G_n}$ for some x and some $n \geq 2$, then $\Gamma = \overline{G_{n+1}}$. To prove this when $n = 2$, one can follow a simple analysis analogous to the proof of Proposition 4. The choices for Δ_x are an edgeless graph on 2 vertices, a 2-regular graph on 4 vertices, or a 3-regular graph on 8 vertices. In the first and last case, the resulting graph Γ is not superregular. In the remaining case, $\Delta_x = C_4$ but there is only one way to join Δ_x to $\Gamma_x \cup \{x\}$ to obtain a superregular graph and the resulting graph is $\overline{G_3}$.

Suppose then that $\Gamma_x = \overline{G_n}$ for $n \geq 3$, and label the vertices of Γ_x with the elements of $\{1, \ldots, n\} \times \{1, \ldots, n\}$ so that

$$(i_1, j_1) \text{ is adjacent to } (i_2, j_2) \text{ if and only if } i_1 \neq i_2 \text{ and } j_1 \neq j_2. \tag{1}$$

The graph induced by neighbors of $(1,1)$ in Γ_x is $\overline{G_{n-1}}$, so $\Gamma_{(1,1)}$ must be $\overline{G_n}$ by the induction hypothesis. Thus $(1,1)$ has exactly $2(n-1)$ neighbors in Δ_x and these may be labeled $(i, 0)$ and $(0, j)$ $(1 \leq i, j \leq n)$ with edges between (i_1, j_1) and (i_2, j_2) according to (1), as long as none of i_1, i_2, j_1, j_2 are 1. Similarly, $\Gamma_{(2,2)}$ must be $\overline{G_n}$, and this requires two more vertices labeled $(1, 0)$ and $(0, 1)$, again with edges according to (1), as long as none of i_1, i_2, j_1, j_2 are 2. Finally label x with $(0, 0)$.

For $k \geq 3$, the neighbors of the vertex (k, k) are now all accounted for. The vertices of $\Gamma_{(k,k)}$ are $\{(i, j) : 0 \leq i, j \leq n \text{ and } i, j \neq k\}$ and the edges are given by (1). There remains to show only that (1) holds for all i_1, i_2, j_1, j_2. If $n \geq 4$ or if $n = 3$ and i_1, i_2, j_1, j_2 are not all different, this follows by consideration of $\Gamma_{(k,k)}$ where k is chosen to be different from all of i_1, i_2, j_1, j_2. If $n = 3$ and i_1, i_2, j_1, j_2 are all different, a count of neighbors of (i_1, j_1) would come up shy of the required 9 unless (i_2, j_2) were included. Thus (1) governs in all situations, and so $\Gamma = \overline{G_{n+1}}$ as claimed.

Lemma 6. *If Γ_x is mK_n for some x, then Γ_x is mK_n for all x.*

Proof. Suppose that $\Gamma_x = mK_n$. Then $\Gamma_y = mK_n$ also for every vertex y in Γ_x, since Γ_y contains K_n as one component, has mn vertices, and is in S. But then $\Gamma_z = mK_n$ for every vertex in Γ, since if $z \in \Delta_x$, then $m \geq 2$ and $z \in \Gamma_y$ for some $y \in \Gamma_x$ (by Proposition 2), and so the same argument applies.

We use Lemma 6 liberally in what follows.

Proposition 7. *For all x, Γ_x is not mK_n.*

Proof. We first address the case $m = 1$. If $\Gamma_x = K_n$, then $\Gamma_y = K_n$ for every vertex y. It follows that $\Gamma = mK_{n+1}$, but this contradicts the fact that $\Gamma \notin S$.

We next address the case $m = 2$. When $n = 2$, this was already addressed at the beginning of the proof of Proposition 5, since $\overline{G_2} = 2K_2$. We proceed with $m = 2$ and $n \geq 3$, proving the stronger result that any superregular graph Γ with $\Gamma_x = 2K_n$ for some x must be G_{n+1}.

Suppose that $\Gamma_x = 2K_n$. Let the two components of Γ_x be denoted A and B and let $A = \{y_1, \ldots, y_n\}$. Then, inside Δ_x, Γ_{y_i} has a component K_n which we denote B_i. The B_i must be disjoint, since $\Gamma_{y_i} = 2K_n$. Now suppose that z_i is a vertex in B_i. By repeatedly applying $\Gamma_z = 2K_n$ for various vertices z, one can infer that z_i has exactly $n-1$ neighbors in $\Delta_x - \Gamma_{y_i}$, none of which are adjacent to any vertices in $\Gamma_{y_i} - \{z_i\}$. Since Δ_x is superregular, the neighbors of z_i induce a $2K_{n-1}$. By the induction hypothesis, $\Delta_x = G_n$. Finally, every vertex in B_i must have exactly one neighbor v in B, and the fact that $\Gamma_v = 2K_n$ forces $\Gamma = G_{n+1}$.

Last, we address the case $m \geq 3$. Assume that $\Gamma_x = mK_n$. We consider the possibilities for Δ_x. Applying Proposition 4 to $\overline{\Gamma}$ rules out $\Delta_x = C_5$ and applying Proposition 5 to $\overline{\Gamma}$ rules out $\Delta_x = G_n$.

Suppose that $\Delta_x = \overline{G_t}$ for some $t \geq 3$ and let y be a vertex in Δ_x, which we now view as $\overline{\Gamma}_x$. Then the neighbors of y in this G_t induce a $2K_{t-1}$, so $\Delta_y = \overline{G_t}$ also. This Δ_y has $(t-1)^2$ vertices in Γ_x and these vertices do not form a clique. But Γ_x has no induced subgraphs whose components are not complete. So $\Delta_x \neq G_t$.

Suppose that $\Delta_x = sK_t$ for some $s, t \geq 1$. Let y be a vertex in Γ_x and z a neighbor of y in Δ_x. Since $\Gamma_y = mK_n$, there are at least $m-1$ neighbors of y in Δ_x that induce a graph with no edges. Thus $s \geq m-1$. Since $\Gamma_z = mK_n$, the neighbors of z in Δ_x must form one copy of K_n, so $t = n+1$. Now the number of edges from Γ_x to Δ_x is $mn^2(m-1)$, while the number of edges from Δ_x to Γ_x is $st(mn-(t-1)) = s(n+1)(mn-n)$. Equating these two counts yields $mn = s(n+1)$, which together with $s \geq m-1$ implies that $s = m-1 = n$. Thus $\Gamma_x = mK_{m-1}$ and $\Delta_x = (m-1)K_m$. But since $m \geq 3$, y has a neighbor y' in Γ_x, and both y and y' have $m-1$ neighbors in each copy of K_m in Δ_x. Hence y and y' have a common neighbor, contradicting the fact that $\Gamma_y = mK_{m-1}$.

Finally, suppose that $\Delta_x = \overline{sK_t}$. The number of edges from Γ_x to Δ_x is $mn^2(m-1)$, while the number from Δ_x to Γ_x is $st(mn-(s-1)t)$. Equating these yields $mn^2(m-1) = st(mn-(s-1)t)$. The fact that $\Gamma_y = mK_n$ and $\Gamma_z = mK_n$ implies that $n \leq s \leq n+1$ and $m-1 \leq t \leq m$, but it is then easy to rule out all the possibilities.

Proposition 8. *For all x, Γ_x is not $\overline{mK_n}$.*

Proof. Since $\overline{mK_1} = K_m$, the case $n = 1$ is handled by Proposition 7. So take $n \geq 2$ and assume that $\Gamma_x = \overline{mK_n}$. An argument similar to Lemma 6 shows that all $\Gamma_y = \overline{mK_n}$. As in Proposition 7, one can then reduce to the case where $\Delta_x = sK_t$ or $\Delta_x = \overline{sK_t}$. The latter case is ruled out by applying Proposition 7 to $\overline{\Gamma}$, so we may assume $\Delta_x = sK_t$ with $t \geq 2$.

If $t = 2$, then the number of edges from Γ_x to Δ_x is $mn(n-1)$ while the number of edges from Δ_x to Γ_x is $2s(mn-1)$. Equating these yields $mn(n-1) = 2s(mn-1)$. This

requires that $2s$ be a multiple of mn, but then the right-hand side becomes larger than the left. If $t \geq 3$, then let y be a vertex in Γ_x and z a neighbor of y in Δ_x. Since $\Gamma_y = \overline{mK_n}$, z must be adjacent to the $m-1$ copies of K_n in $\overline{\Gamma_x - \Gamma_y}$. Since $\Gamma_z = \overline{mK_n}$, the remaining neighbors of z must induce a subgraph with no edges. But since $t \geq 3$, z has a pair of adjacent neighbors in Δ_x, a contradiction.

Proposition 9. *For all x, Γ_x is not G_n.*

Proof. Since $\overline{G_3} = G_3$, we may assume $n \geq 4$. Assume that $\Gamma_x = G_n$. Then Γ is n^2-regular, and every vertex in Γ_x has exactly $(n-1)^2$ neighbors in Δ_x. So the number of edges from Γ_x to Δ_x is $n^2(n-1)^2$. Also, since Δ_x must be $\overline{G_t}$ for some $t \geq 4$, every vertex in Δ_x has exactly $n^2-(t-1)^2$ neighbors in Γ_x, so the number of edges from Δ_x to Γ_x is $t^2(n^2-(t-1)^2)$. Setting these equal yields

$$\left(\frac{n-1}{t}\right)^2 + \left(\frac{t-1}{n}\right)^2 = 1.$$

It is easy to see that this has no solutions in integers greater than 2.

Editorial comment. Five claimed solutions to this problem were available when the deadline for solutions was reached, but none turned out to be correct. Later, Douglas B. West solved the problem, but as a report on his solution was being prepared, a solution from Randall B. Maddox was received. What we have presented above is a digest of Maddox's proof. In order to illustrate the organization of the solution in limited space, many details are left to the reader. West's solution has been accepted for publication in *J. Graph Theory*.

A Square Crossing

10322 [1993, 688]. *Proposed by Jiang Huanxin, student, FuDan University, ShangHai, China.*

Let $ABCD$ and $AEFG$ be squares with the common vertex A and *different* edge lengths. Let $\theta = \angle EAD$ $(0 < \theta < \pi/2)$. Suppose that EF and CD intersect at the point P. For which value of θ will $AP \perp CF$?

Solution by H. Sunil Gunaratne, Universiti Brunei Darussalam, Gadong, Brunei. Assume $|AE| : |AD| = \lambda : 1$, with $\lambda > 0$ and $\lambda \neq 1$. Then there are two cases. **Case (i)**: $AEFG$ has the same orientation as $ABCD$. Then $\theta = \pi/4$ is the unique solution, independent of λ. **Case (ii)**: $AEFG$ has the orientation opposite to that of $ABCD$. Then the solutions are of the form $\theta = \alpha \pm \beta$ where $-\pi/2 < \alpha, \beta < \pi/2$ with $\cos\alpha = (\lambda^2+1)(2+\lambda^4)^{-1/2}$, $\cos\beta = 2\lambda(2+\lambda^4)^{-1/2}$. In addition, α and β should have the same sign, with $0 < \alpha < \pi/2$ if $\lambda > 1$ and $-\pi/2 < \alpha < 0$ if $0 < \lambda < 1$. It is also necessary that $\sqrt{2}-1 < \lambda < \sqrt{2}+1$ in order to have $-\pi/2 < \theta < \pi/2$.

To show this, assume without loss of generality that $|AD| = 1$, $|AE| = \lambda$, and use vectors based at A. Thus, we write $\overrightarrow{AB} = \mathbf{i}$, $\overrightarrow{AD} = \mathbf{j}$, $\overrightarrow{AE} = \lambda\mathbf{e}$, $\overrightarrow{AG} = \lambda\mathbf{g}$, where $\mathbf{i}, \mathbf{j}, \mathbf{e}, \mathbf{g}$ are unit vectors and $\mathbf{i}\cdot\mathbf{j} = 0$. Then we have

$$\mathbf{e} = \mathbf{i}\sin\theta + \mathbf{j}\cos\theta$$
$$\pm\mathbf{g} = -\mathbf{i}\cos\theta + \mathbf{j}\sin\theta$$

with $-\pi/2 < \theta < \pi/2, \theta \neq 0$. The plus sign is taken in Case (i) and the minus sign in Case (ii). Then $\overrightarrow{AP} = \overrightarrow{AD} + \overrightarrow{DP} = \mathbf{j} + t\mathbf{i} = \overrightarrow{AE} + \overrightarrow{EP} = \lambda\mathbf{e} + r\lambda\mathbf{g}$ for scalars r and t to be

determined. Taking the dot product with $\mathbf{e}$ yields $\mathbf{j}\cdot\mathbf{e}+t\mathbf{i}\cdot\mathbf{e}=\lambda$, or $t=(\lambda-\cos\theta)/\sin\theta$ (note that $\theta\neq 0$). Now $AP\perp CF$ if and only if

$$\begin{aligned}0=\overrightarrow{AP}\cdot\overrightarrow{CF}&=\left(\mathbf{j}+t\mathbf{i}\right)\cdot\left(\lambda\mathbf{e}+\lambda\mathbf{g}-\mathbf{i}-\mathbf{j}\right)\\&=\lambda\cos\theta-1+t\left(\lambda\sin\theta-1\right)\pm\lambda\left(\sin\theta-t\cos\theta\right).\end{aligned}\tag{1}$$

Substituting for t in terms of λ in (1) and simplifying yields the necessary and sufficient condition

$$\left(\lambda^2-1\right)\sin\theta-\left(\pm\lambda^2-1\right)\cos\theta\pm\lambda-\lambda=0.\tag{2}$$

In **Case (i)**, the plus sign in (2) gives

$$\left(\lambda^2-1\right)\left(\sin\theta-\cos\theta\right)=0.\tag{3}$$

Since $\lambda\neq 1$, (3) gives $\sin\theta=\cos\theta$, and therefore $\theta=\pi/4$ is the only solution since $-\pi/2<\theta<\pi/2$. Note that if $\lambda=1$, then $AP\perp CF$ for all θ with $-\pi/2<\theta<\pi/2$ and $\theta\neq 0$.

In **Case (ii)**, the minus sign in (2) gives

$$\left(\lambda^2+1\right)\cos\theta+\left(\lambda^2-1\right)\sin\theta=2\lambda\tag{4}$$

This equation reduces to $\cos(\theta-\alpha)=\cos\beta$ with α and β defined as at the start of this solution. Since α, β and θ all lie between $-\pi/2$ and $\pi/2$, this requires that $\theta=\alpha\pm\beta$. It can be directly verified that this value of θ satisfies (4) and that there are at most two values of θ satisfying (4) and the given constraints. Note that $-\pi/2<\alpha-\beta<\pi/2$ since α and β were chosen in the same quadrant, while $-\pi/2<\alpha-\beta<\pi/2$ if and only if

$$0<\cos(\alpha+\beta)=\frac{\sqrt{2}\lambda\left(\lambda^2+1\right)-\left(\lambda^2-1\right)^2}{\left(\lambda^4+1\right)\sqrt{2}}.$$

Since $\sqrt{2}\lambda\left(\lambda^2+1\right)-\left(\lambda^2-1\right)^2=\left(1+\sqrt{2}-\lambda\right)\left(\lambda-\sqrt{2}+1\right)\left(\lambda^2-\sqrt{2}\lambda+1\right)$, we have $\cos(\alpha+\beta)>0$ if and only if $\sqrt{2}-1<\lambda<\sqrt{2}+1$.

Editorial comment. The selected solution was the only one to deal with the case in which the given squares have opposite orientation. The proposer included a figure indicating that the squares were intended to be similarly oriented, and some solvers noted that they assumed that the vertices of each figure were listed in counterclockwise order. Most solutions used coordinates, vectors, or complex numbers leading to the equivalent of (3). The solutions of the proposer, Adam Coffman, Richard Holzsager, and (a joint solution by) C. W. Wample & W. W. Meyer employed purely synthetic methods. The key to such a solution is that if A and P are fixed, the locus of possible choices for C is a circle.

Solved by 41 readers and the proposer. There were four incorrect or incomplete solutions.

A Divisibility Property of the Central Binomial Coefficient

10326 [1993, 689]. *Proposed by Ira Gessel, Brandeis University, Waltham, MA.*

For r a positive integer, let K_r be the smallest positive integer such that

$$\frac{K_r}{n+r}\binom{2n}{n}$$

is an integer for all $n\geq 0$. Show that

$$K_r=\frac{r}{2}\binom{2r}{r}.$$

Solution by A. N. 't Woord, University of Technology, Eindhoven, the Netherlands. We first show that $K_r \leq \frac{r}{2}\binom{2r}{r}$ by proving that

$$A(n,r) = \frac{r}{2(n+r)}\binom{2n}{n}\binom{2r}{r}$$

is an integer for all $n \geq 0$ and $r \geq 1$. Note that

$$A(0,r) = \binom{2r-1}{r} \text{ and } A(n,1) = \frac{1}{n+1}\binom{2n}{n},$$

the latter being a Catalan number, so $A(0,r)$ and $A(n,1)$ are integers for all n and r. Since

$$\binom{2m+2}{m+1} = 2\binom{2m+1}{m} = 2\frac{2m+1}{m+1}\binom{2m}{m},$$

we have

$$\begin{aligned}
A(n,r+1) - A(n+1,r) &= \\
&= \frac{r+1}{2(n+r+1)}\binom{2n}{n}\binom{2r+2}{r+1} - \frac{r}{2(n+r+1)}\binom{2n+2}{n+1}\binom{2r}{r} \\
&= \frac{1}{n+r+1}\binom{2n}{n}\binom{2r}{r}\left(2r+1-r\frac{2n+1}{n+1}\right) \\
&= \frac{1}{n+1}\binom{2n}{n}\binom{2r}{r} = 2A(n,1)A(0,r).
\end{aligned}$$

By induction on r, therefore, $A(n,r)$ is an integer for all n and r.

Suppose now that, for some r, $K_r < L_r = \frac{r}{2}\binom{2r}{r}$. The integer K_r is the least common multiple, over all $n \geq 0$, of the denominators of the numbers $\frac{1}{n+r}\binom{2n}{n}$ when written in lowest terms. By the argument above, L_r is a multiple of all these denominators. Hence K_r divides L_r. By the definition of K_r, any prime p that divides L_r/K_r also divides $A(n,r)$ for each $n \geq 0$. Since $2A(n,1)A(0,s) + A(n+1,s) = A(n,s+1)$, we have by induction on m that p divides $A(n,m)$ for all $n \geq 0$ and $m \geq r$.

Now choose k such that $p^k \geq r$. Since p divides $\binom{p^k}{i}$ for $1 \leq i \leq p^k - 1$, the identity $\binom{2n}{n} = \sum_{i=0}^{n}\binom{n}{i}^2$ yields $\binom{2p^k}{p^k} \equiv 2 \pmod{p}$. Therefore p does not divide $\frac{1}{2}\binom{2p^k}{p^k} = A(0,p^k)$. This contradicts the preceding paragraph, so $K_r = \frac{r}{2}\binom{2r}{r}$ for all r.

Solved also by D. Callan, R. J. Chapman (U. K.), R. Holzsager, R. Richberg (Germany), Anchorage Math Solutions Group, and the proposer.

A Continued Fraction

10327 [1993, 689]. *Proposed by Jerome Minkus, Berkeley, CA.*

Find the simple continued fraction for $(e+3)/4$.

Solution by Jean Anglesio, Garches, France. We prove that

$$\frac{e+3}{4} = 1 + \frac{1}{2+}\,\frac{1}{3+}\,\frac{1}{20+}\cdots\frac{1}{(4n-1)+}\,\frac{1}{(16n+4)+}\cdots.$$

We start from the simple continued fraction for $(e-1)/2$ attributed to Euler (1737), which is $(e-1)/2 = 0 + \frac{1}{1+}\,\frac{1}{6+}\cdots\frac{1}{(4n+2)+}\cdots$ (see for example O. Perron, *Die Lehre von den Kettenbrüchen*, Chelsea, 1929, p124). We use the concise notation $(e-1)/2 = [0;1,6,10,\ldots,4n+2,\ldots] = [0;1,\overline{4n+2}]_{n=1}^{\infty}$. Since

$$\frac{1}{2}\left(\frac{1}{a_1+}\,\frac{1}{a_2+}\cdots\frac{1}{a_n+}\cdots\right) = \frac{1}{2a_1+}\,\frac{1}{(a_2/2)+}\cdots\frac{1}{2^{(-1)^n}a_n+}\cdots,$$

we have

$$\frac{e+3}{4} = 1 + \frac{1}{2}\frac{e-1}{2} = 1 + \frac{1}{2}[0; 1, \overline{8n-2, 8n+2}]_{n=1}^{\infty} = [1; 2, \overline{4n-1, 16n+4}]_{n=1}^{\infty}.$$

Editorial comment. The proposer proved more generally that $\left(e^{1/k} + 4k - 1\right) / \left(4k\right) = [1, 4k^2 - 2k, \overline{4n-1, 4k^2(4n+1)}]_{n=1}^{\infty}$. The alternating pattern of partial quotients can be traced to the fact that all partial quotients of $\left(e^{1/k}+1\right) / \left(e^{1/k}-1\right)$ are even. Related continued fractions can be found in K. R. Matthews and R. F. C. Walters, "Some properties of the continued fraction expansion of $(m/n)e^{1/a}$", *Proc. Camb. Phil. Soc.* 67(1970), 67–74. István Nemes based his method of solution on a process for finding the continued fraction of 2α from that of α that can be found in D. E. Knuth, *The Art of Computer Programming, Vol II: Seminumerical Algorithms*, Addison-Wesley, 1969, section 4.5.3, exercise 14, where it is attributed to Hurwitz. More general algorithms are also available. If α is any number whose continued fraction is known, and T is any linear fractional transformation, the method of G. N. Raney, "On continued fractions and finite automata", *Math. Annalen* 206 (1973), 265–283 allows the continued fraction of $T(\alpha)$ to be found.

Solved also by B. D. Beasley, J. Boutillon (France), R. J. Chapman (U. K.), Z. Franco, C. Georghiou (Greece), R. Holzsager, D. J. Jones, M. J. Knight, A. Kulik, D. C. Kurtz, I. Nemes (Austria), R. M. Robinson, H.-J. Seiffert (Germany), N. C. Singer, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, and the proposer. Two incorrect solutions were received.

Derived Matrices with the Same Characteristic Polynomial

10334 [1993, 797]. *Proposed by John Sarli, California State University, San Bernardino, CA.*

Let M be a fixed n by n matrix with complex entries which is *not* nilpotent. For $a, b \in \mathbb{C}$, define the linear operator $M_{a,b}$ on the space of n by n complex matrices by $M_{a,b}(N) = aMN + bNM$. If the operators $M_{a,b}$ and $M_{c,d}$ have the same characteristic polynomial, show that $a^k + b^k = c^k + d^k$ for some k with $1 \le k \le n$.

Solution by David Callan, University of Wisconsin, Madison, WI. Let $U = \{\lambda_i\}_{i=1}^n$ denote the spectrum of M, and let $U_{a,b}$ denote the spectrum of $M_{a,b}$. It is well known from the theory of Kronecker products (see Roger A. Horn & Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge, 1991, pp. 268–269) that $U_{a,b} = \left\{a\lambda_i + b\lambda_j : 1 \le i, j \le n\right\}$. The hypothesis is that $U_{a,b} = U_{c,d}$. Equating the sums of the elements of $U_{a,b}$ and $U_{c,d}$ yields $a + b = c + d$ unless $\sum \lambda_i = 0$. In this case, equating the sums of their squares yields $a^2 + b^2 = c^2 + d^2$ unless also $\sum \lambda_i^2 = 0$. Continuing in this way yields the assertion of the proposal, since $\sum \lambda_i^k$ cannot equal 0 for all $1 \le k \le n$ unless M is nilpotent.

Solved also by M. Marcus, A. Tissier (France), P. Y. Wu (Canada), and the proposer.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

Lion Hunting and Other Mathematical Pursuits. A Collection of Mathematics, Verse, and Stories by Ralph P. Boas, Jr. Edited by Gerald L. Alexanderson and Dale H. Mugler, The Mathematical Association of America, 1995, xii + 308 pp., $35.00.

Reviewed by **Leonard Gillman**

Loud cheers for Donald Albers, MAA Director of Publications, for suggesting this book, and for Gerald Alexanderson and Dale Mugler for creating it!

The late Ralph Boas was a mathematician of diverse interests and accomplishments: research mathematician (close to 200 research papers, several with very distinguished co-authors such as S. Bochner, D. V. Wider, G. Pólya, N. Levinson, M. Kac, P. Erdös, K. Chandresakharan); teacher; writer (*Entire functions*, 1954, *A primer of real functions*, 1960, *Invitation to complex analysis*, 1987, as well as verses and fiction); editor (*Mathematical Reviews*, *American Mathematical Monthly*, "a position secretly coveted"); and department chairman (Northwestern University, 15 years). This book collects a number of his lighter writings in and outside mathematics, the latter including reminiscences and anecdotes as well as limericks, clerihews, and other verses (but not *The Versed of Boas*, 1983).

The stage is set by a twenty-page autobiographical sketch (taken from *More Mathematical People*, D. J. Albers et al., Harcourt Brace Jovanovich, 1990). It is a fascinating tour through Boas's school days, college years, and professional life, and includes a charming account of how he met his wife, Mary, as well as photographs of his family and of some of the eminent mathematicians he encountered.

The book proper presents enough of a mix to delight readers of many stripes: reprints of a good number of Boas' expository mathematical papers (on Lion Hunting, Infinite Series, The Mean-Value Theorem and Indeterminate Forms, Complex Variables, Inverse Functions, Polynomials, The Teaching of Mathematics); several sections titled Recollections and Verse; Reminiscences by several close friends, students, and son Harold (also a mathematician); Reviews and other miscellany.

What I found most engaging were the autobiography and other sections containing Boas' insightful comments about mathematics or about people and their foibles. His prose is simple and direct, and at the same time, elegant and witty:

From Boas's teaching stint at the Navy Pre-Flight School in Chapel Hill during World War II:

> One cadet, who had a private airplane pilot's license, was failing mathematics. When he was asked how much gas he would need to carry if he were going to fly two hundred miles at so many miles per gallon, he didn't know

whether to multiply or divide. How, the officers asked, was he able to get the right answer? He replied that he did it both ways and took the reasonable answer. They felt that anybody who knew what was a reasonable answer had promise, so they gave him a second chance.

From Boas' year in Princeton:

> Bochner had a number of standard responses to any problem you asked him about. They ranged from "I think this is not very interesting" to "I think this cannot be." Once I got "I think this is difficult" and then solved the problem. When I took the result to Bochner, he said, "I think this is trivial."

From Boas' years as department chairman:

> I received a deputation of students who complained about their instructor. They were followed by another group of students who... wanted me to know what a wonderful teacher they had. Same class, same teacher.

From Boas' years as Editor of *Mathematical Reviews*:

> I have always been charmed by the fact that the Georgian word for "father" is "mama". This controverts psychologists who have convincing reasons why a child calls its mother "mama". (In case you are wondering, the Georgian word for "mother" is "dedi".)

The text proper starts off with several papers on the theory of big game hunting. The first of them, written by Ralph Boas and Frank Smithies under the pseudonym H. Petard, itself a pseudonym for E. S. Pondiczery, which was in turn a Slavic rendition of Pondicherry (a name they happened to be fond of), appeared in the *Monthly* **45** (1938) 446–447; "Pondiczery" had sent a cover letter explaining that he would prefer to have the paper published pseudonymously. The existence of "E. S. Pondiczery, New York City" had been carefully established by means of an earlier note in the same volume (p. 307), *A function-theoretic paradox* (which showed that all analytic functions are constant). The paper on lion hunting stemmed from dinner conversations in Princeton about the lion-hunting methods devised in Göttingen. The name "H. Petard", by the way, comes from Hamlet's lines (III, iv, 206–7):

For 'tis the sport to have the enginer
Hoist with his own petar.

Without loss of generality, Dr. Pondiczery confines his attention to Lions (*Felis leo*) of the Sahara Desert. My favorites are methods 1 and 11:

1. THE HILBERT, OR AXIOMATIC METHOD. We place a locked cage at a given point of the desert. We then introduce the following logical system.

Axiom I. The class of lions in the Sahara Desert is non-void.

Axiom II. If there is a lion in the Sahara Desert, there is a lion in the cage.

Rule of Procedure. If p is a theorem and "p implies q" is a theorem, then q is a theorem.

Theorem 1. *There is a lion in the cage.*

11. THE SCHRÖDINGER METHOD. At any given moment there is a positive probability that there is a lion in the cage. Sit down and wait.

The paper proved to be "seminal": it spawned a healthy series of sequels, of which a half dozen are reproduced in the book. Some of the methods are intriguing, but after a while the joke begins to pall.

Of the serious mathematical papers, I found the most interesting to be the discussion of the Mean-Value Theorem and Indeterminate Forms, and the papers on Teaching. Following Dieudonné, Boas espouses the Mean-Value Inequality as a replacement for the Theorem, asserting that it is more intuitive, and pointing out that you can still derive the usual results and in fact presenting a simple proof of L'Hopital's Rule (including the case "∞/∞"). (I have some questions about his derivation, but this is not the place to go into them.)

As for Boas' verses, many of them are clever and well done (I particularly like Echolalia), but several are marred by couplets whose lines have varying length or meter, or do their rhyming on unstressed syllables:

Donald Saari:
His eyes are starry
From applying celestial mechanics
To representational politics.

I thought the fiction pieces (all of which were new to me) were of high quality. Two are outstanding: *The Rose Acacia*, concerning a Faustian contract between a computer and Satan, and *Game Adjourned*, concerning the female mathematical genius Valentina Pondiczery, who works out a winning strategy for White in chess.

The papers on teaching, which bear provocative titles such as *If This Be Treason*... (1957), *Calculus as an Experimental Science* (1971), *Can We Make Mathematics Intelligible*? (1981), are thought-provoking and full of good common sense. I close this review with the following "treasonable" quotation (slightly compressed):

> I claim that there is a place, and a use, even for the solution of quartics by radicals, or Horner's method, or involutes and evolutes, or whatever your particular candidates for oblivion may be. Here are problems that might conceivably have to be solved; perhaps the methods are not the most practical ones; but that is not the point. The point is that in solving the problems the student gets practice in using the necessary mathematical tools, and gets it by doing something that has more motivation than mere drill.
>
> It is the fashion to deprecate puzzle problems and artificial story problems. I think that there is a place for them too. Problems about mixing chemicals or sharing work, however unrealistic, give good practice and even have a good deal of popular appeal. It is absurd to claim that only "real" applications should be used to illustrate mathematical principles. Most of the real applications are too difficult and/or involve too many side issues. One begins the study of French with simple artificial sentences, not with the philosophical writings of M. Sartre.
>
> The traditional topics have persisted partly by mere inertia but partly because they still serve a real purpose, even if it is not their ostensible purpose. Let us keep this in mind when we are revising the curriculum.

1606 The High Road, Austin TX 78746
gillman@mail.utexas.edu

Nature's Numbers. By Ian Stewart, New York, Basic Books, 1995

Reviewed by **Freeman J. Dyson**

This book describes in clear and simple language a view of mathematics with which I profoundly disagree. It gives me a good opportunity to say why I disagree, to put his view and mine into historical perspective. Stewart's book is about the new wave in mathematics that grew around the ideas of non-linear dynamics, chaos theory, and complexity theory. The new wave began with classical physics, especially the physics of fluids. It flourished and became popular when computers were able to simulate classical motions accurately and display them dramatically. Stewart emphasizes the fact that the style of the new mathematics is visual rather than analytical. He ends his book with a prediction, that a new discipline which he calls "morphomatics" will give us deep understanding of natural patterns, especially the patterns of biological growth and form.

I have no quarrel with any of Stewart's positive statements. He has every right to be enthusiastic about the new visual style of mathematical thinking. The visual style has made mathematics more accessible to the general public and more concerned with things that the general public can understand. Pictures of clouds or water-droplets bring mathematics to millions whom differential equations will never reach. I disagree with Stewart when he makes negative statements about old-fashioned non-visual mathematics, the mathematics of equations and exact solutions. He is especially negative about quantum mechanics. He hopes that quantum mechanics will one day turn out to be an effect of classical chaos working in some undiscovered realm of hidden variables. He is blind to the beauty of quantum mechanics, a beauty that is to me transcendent and overwhelming. The idea of explaining quantum mechanics with classical models is to me as absurd as the efforts of James Clerk Maxwell to explain his electromagnetic theory with mechanical models. The beauty of Maxwell's equations becomes visible only when you abandon mechanical models, and the beauty of quantum mechanics becomes visible only when you abandon classical thinking.

It is easy to see why Stewart dislikes quantum mechanics. Quantum mechanics runs counter to the two cardinal principles of the new wave of mathematics. Quantum mechanics is precisely linear, and profoundly non-visual. The linearity of quantum mechanics leads to an amazing richness of exact mathematical structures and exact solutions. The abstract nature of quantum concepts such as particle spins or anticommuting fields makes them impossible to visualize. There is no way to represent the quantum world by a picture on a computer screen. A picture is a good tool for describing a classical fluid, but it cannot describe a quantum fluid such as liquid helium. The beauty of quantum mechanics is an abstract beauty. The symmetries of quantum systems are described in the abstract language of Lie Algebras rather than in the visual language of geometry. The dynamics of quantum systems cannot be simulated by motions in our familiar three-dimensional world. The space of quantum mechanics is Hilbert space, a space with infinitely many dimensions. Einstein said in his triumphant presentation of the theory of general relativity to the Prussian Academy of Sciences in Berlin in 1915, "Hardly anybody who has really grasped the theory will be able to escape from its magic". In my view, this statement is as true of quantum mechanics as of general relativity. Once

you have experienced the vastness of Hilbert space and the power of functional analysis, you do not want to go back to the small world of classical mechanics and three-dimensional geometry. But Stewart has not really grasped quantum mechanics and is immune to its magic.

Stewart and I belong to different generations. When I was a student fifty years ago, the new wave in mathematics was abstract algebra, the new wave in physics was quantum mechanics. As a student, with the normal arrogance of youth, I considered the new stuff to be the only stuff worth knowing and despised everything that had gone before. I took pride in not reading any papers that were more than five years old. Quantum mechanics was elegant and powerful. Classical physics was ugly and complicated, not worth the time it would take to learn it. So I jumped over classical physics and made a career doing the fashionable stuff, understanding the quantum theory of electromagnetic processes. Now, fifty years later, the fashions have come and gone, the fashionable has become unfashionable and the unfashionable has become fashionable. Stewart loves hydrodynamics and despises quantum mechanics, just as I loved quantum mechanics and despised hydrodynamics fifty years ago. Stewart is now fashionable and I am unfashionable. That is as it should be. Changing fashions are an essential part of the growth of mathematics. I do not regret the change of fashion. I only wish to point out the impermanence of all fashions and the importance of keeping unfashionable areas of mathematics alive.

In 1942, when I was a student in Cambridge and quantum mechanics was riding high, I heard a lecture by Mary Cartwright about the Van der Pol equation. Cartwright had been working with Littlewood on the solutions of the equation, which describe the output of a non-linear radio amplifier when the input is a pure sine-wave. The whole development of radar in World War Two depended on high-power amplifiers, and it was a matter of life and death to have amplifiers that did what they were supposed to do. The soldiers were plagued with amplifiers that misbehaved, and blamed the manufacturers for their erratic behavior. Cartwright and Littlewood discovered that the manufacturers were not to blame. The equation itself was to blame. They discovered that as you raise the gain of the amplifier, the solutions of the equation become more and more irregular. At low power the solution has the same period as the input, but as the power increases you see solutions with double the period, then with quadrupole the period, and finally you have solutions that are not periodic at all. Cartwright and Littlewood explored the behavior of the solutions in detail and discovered the phenomena that later became known as "chaos". They published all this in a paper in the *Journal of the London Mathematical Society*, which appeared in 1945. That was a bad time to publish. Paper in England was scarce and few copies of the journal were printed. Mathematicians everywhere were still busy fighting the war. The paper attracted no attention. In 1949 Mary Cartwright came to Princeton and talked about the work again. Again she attracted no attention. Littlewood was not helpful. In the foreword to Littlewood's collected papers is a description written by Littlewood about his collaboration with Cartwright:

"Two rats fell into a can of milk. After swimming for a time one of them realized his hopeless fate and drowned. The other persisted, and at last the milk was turned to butter and he could get out."

Littlewood does not say whether the rat who drowned was himself or Cartwright. In either case, the passage makes clear that Littlewood did not understand the importance of the work that he and Cartwright had done. Only Cartwright understood it, and she is not a person who likes to blow her own trumpet. She put

the Van der Pol equation aside and went on to a distinguished career in analytic function theory and university administration. She became President of the London Mathematical Society in 1961, and Dame Mary (the female equivalent of a knighthood) in 1969. By that time, the phenomena of chaos had been rediscovered by Edward Lorenz. A few years later, they were given their modern names.

I tell this story of the Van der Pol equation because it illustrates vividly the blindness of mathematicians to discoveries in unfashionable fields. I was myself as blind as everybody else to the importance of Cartwright's work. When I heard her lecture in 1942, I remember being delighted with the beauty of her results. She had disentangled what Littlewood himself called "the dramatic fine structure of solutions". I could see the beauty of her work but I could not see its importance. I said to myself, "This is a lovely piece of work. Too bad it is only a practical wartime problem and not real mathematics." I did not say, "This is the beginning of a new era, a new way of doing mathematics." I did not say, "This is the birth of a new field of mathematics that I could spend the rest of my life exploring." I missed an opportunity to be the pioneer of a new field, because I shared the tastes and prejudices of my contemporaries.

The disagreement between Stewart and me is mainly a matter of taste. My mathematical tastes were formed at a time when the most common adjectives in mathematical conversation were "deep" and "trivial". First-rate mathematics was deep, all the rest was trivial. Logical depth was the essential virtue that we looked for in mathematics. Stewart is not interested in logical depth. His book contains many examples of mathematical description, but not one that I would consider deep. For him the language of mathematics is form and pattern that can be visually displayed. The essential virtue of mathematics is to display patterns that appear in a wide variety of different circumstances. His examples are well chosen to illustrate the power of mathematics to bring processes from population biology and embryology and meteorology into a unified picture. For me, an old-fashioned analyst, these examples are visually interesting but mathematically trivial. I miss the depth that complex analysis brought to number theory, the depth that Lie algebras and fibre-bundles brought to physics.

I see some parallels between the shifts of fashion in mathematics and in music. In music, the new popular styles of jazz and rock became fashionable a little earlier than the new mathematical styles of chaos and complexity theory. Jazz and rock were long despised by classical musicians, but have emerged as art-forms more accessible than classical music to a wide section of the public. Jazz and rock are no longer to be despised as passing fads. Neither are chaos and complexity theory. But still, classical music and classical mathematics are not dead. Mozart lives, and so does Euler. When the wheel of fashion turns once more, quantum mechanics and hard analysis will once again be in style.

Institute for Advanced Study
Princeton, New Jersey

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Qucstionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, S, L. *Keys to Infinity.* Clifford A. Pickover. Wiley, 1995, xviii + 331 pp, $24.95. [ISBN 0-471-11857-5] A delightful, variegated, yet turbulent anthology of patterns that emerge from recursion, limits, randomness, and other infinite processes. Employs essays, puzzles, computer code, images, and quotations to explore a mathematical phantasmagoria including such unheralded creations as batrachions and mygalomorphs. A wonderful grab-bag of tricks with which to dazzle the mathematically literate. LAS

General, T(13–14: 2), S. *Brückenkurs Mathematik für Wirtschaftswissenschaftler.* Walter Purkert. BG Teubner Leipzig, 1995, 436 pp, DM 44 (P). [ISBN 3-8154-2080-6] On the secondary-school mathematics German university economics students need to know. Treats arithmetic of reals; powers, roots, logarithms; sequences and series; calculus; linear algebra. Exercises and solutions, no proofs. JD-B

Precalculus, T*(13: 1). *Precalculus Concepts in Context.* Judith Flagg Moran, Marsha Davis, Mary Murphy. PWS, 1996, xvii + 555 pp, $60.50 (P). [ISBN 0-534-92157-4] An interactive text that engages students with write-in questions, collaborative activities, expository writing assignments, labs, projects. Emphasizes functions, graphing technology. TH

Education, S(15–18). *Connecting Mathematics across the Curriculum: 1995 Yearbook.* Peggy A. House, Arthur F. Coxford. NCTM, 1995, viii + 245 pp, $20. [ISBN 0-87353-394-1] Issues, ideas, and examples for helping students make connections among mathematical ideas, between mathematics and other school subjects, and between mathematics and real-world contexts. Lead article discusses connections as (1) unifying themes (e.g., change, shape); (2) mathematical processes (e.g., representation, problem solving); (3) mathematical connectors (e.g., functions, variable). Other "issue" articles make connections with school science, ethnomathematics, and history. Specific ideas for elementary, middle, and high school students. Excellent resource for pre- and in-service teachers. MW

Education, S(15–18). *Algebra in a Technological World.* M. Kathleen Heid, *et al.* Curr. & Eval. Standards for School Math. Addenda Ser., Grades 9–12. NCTM, 1995, viii + 168 pp, $15 (P). [ISBN 0-87353-326-7] Explores the effects of graphing technology and computer algebra systems on the content, focus, and emphasis of school algebra. Includes a chapter on a modeling and functions approach to algebra, and one on connections among algebra, geometry, and discrete mathematics. Eighteen complete student activities on black-line masters, and numerous examples and teaching suggestions illuminate necessary and possible changes in learning and teaching. MW

Education, T(17–18: 1), S, P. *Mathematics Education: Models and Processes.* Lyn D. English, Graeme S. Halford. Lawrence Erlbaum Assoc, 1995, xii + 360 pp, $34.50 (P); $69.95. [ISBN 0-8058-1458-2; 0-8058-1457-4] Theoretical look at the learning of elementary computation, algebra, proportional reasoning, and mathematical problem solving from a cognitive

science perspective. Traces history of the psychology of mathematics education, and summarizes principles of cognition and cognitive development. Complete bibliography. Valuable reference for mathematics educators and graduate students. MW

Education, S(17–18), P. *New Directions for Equity in Mathematics Education.* Eds: Walter G. Secada, Elizabeth Fennema, Lisa Byrd Adajian. Cambridge Univ Pr, 1995, xi + 364 pp, $16.95 (P); $49.95. [ISBN 0-521-47720-4; 0-521-47152-4] Scholarship on the nature of equity that incorporates alternative conceptions of equity and new methods of inquiry. Project descriptions, research reports, and policy discussions examine the role that mathematics education plays in social stratification and could play in enabling "a fairer social order." MW

Education, S(17–18), P. *Equity in Mathematics Education: Influences of Feminism and Culture.* Eds: Pat Rogers, Gabriele Kaiser. Falmer Pr, 1995, ix + 278 pp, $24.95 (P). [ISBN 0-7507-0401-2; 0-7507-0400-4] Overview of current thinking grounded in the five phases of the McIntosh model for interactive curricular and personal revision: (1) womanless mathematics; (2) (a few) women in mathematics; (3) women as a problem in mathematics; (4) women as central to mathematics; and (5) mathematics reconstructed to include all. Must reading for those seeking a feminist perspective. MW

Logic, P. *The Logical Status of Diagrams.* Sun-Joo Shin. Cambridge Univ Pr, 1994, xi + 197 pp. [ISBN 0-521-46157-X] Elaborates a case study of Venn-type diagrams. Provides a careful syntax, semantics, and set of transformation rules, and proves they form a sound and complete formal representation system equivalent to a first-order language with one place predicates, existential, and universal quantifiers. Insightful discussion of diagrammatic vs. linguistic representation systems. RM

Logic, P. *Algebraic Set Theory.* A. Joyal, I. Moerdijk. London Math. Soc. Lect. Note Ser., V. 220. Cambridge Univ Pr, 1995, viii + 123 pp, $34.95 (P). [ISBN 0-521-55830-1] Algebraic approach to set theory encompassing both Zermelo-Fraenkel set theory and topos theory. LB

Algebra, P. *Algebraic K-Theory.* Hvedri Inassaridze. Math. & Its Applic., V. 311. Kluwer Academic, 1995, viii + 438 pp, $189. [ISBN 0-7923-3185-0] Principal algebraic K-theories; connections with topological K-theory, cyclic homology, applications to monoid, and polynomial algebras. TH

Real Analysis, T(16: 1), L. *Measure Theory and Probability.* Malcolm Adams, Victor Guillemin. Birkhäuser Boston, 1996, xiv + 205 pp, $26.50. [ISBN 0-8176-3884-9] Reprint, with corrections, of 1986 Wadsworth edition (TR, December 1986). LC

Real Analysis, T(14–15). *Introduction to Real Analysis.* Michael J. Schramm. Prentice Hall, 1996, xiii + 368 pp. [ISBN 0-13-229824-4] First half is an introduction to proof techniques and the completeness of the real numbers motivated by examining the similarities and differences between the rational numbers and the reals. Second half covers the standard topics from calculus. TR

Differential Equations, T(14–15: 1). *Introduction to Differential Equations with Boundary Value Problems.* Stephen L. Campbell, Richard Haberman. Houghton Mifflin, 1996, xiii + 737 pp, [ISBN 0-395-70828-1]; *Solutions Manual*, 178 pp, (P). [ISBN 0-395-74632-9] Fairly traditional text. Chapters on discrete dynamical systems, Fourier series, partial differential equations. LC

Operator Theory, T(18), S, P. *Introduction to Spectral Theory With Applications to Schrödinger Operators.* P.D. Hislop, I.M. Sigal. Appl. Math. Sci., V. 113. Springer-Verlag, 1996, ix + 337 pp, $49. [ISBN 0-387-94501-6] Written with the philosophy that the heart of modern analysis lies in its application to other disciplines. Uses modern physics (Schrödinger operators, quantum tunneling, and resonance) to motivate a self-contained study of the spectral analysis of linear differential operators. Emphasizes geometric methods. Knowledge of Banach spaces required. SA

Operator Theory, T(18), S. *Fourier Integral Operators.* J.J. Duistermaat. Progress in Math., V. 130. Birkhäuser Boston, 1996, viii + 142 pp, $38.50. [ISBN 0-8176-3821-0] Lecture notes for a course given in 1970. An introduction, not aimed at recent applications. This powerful technique for solving PDE's requires some differential geometry, and this subject is given thorough treatment. (1973 paperback, TR, March 1974.) SA

Analysis, T(15), S. *Analysis by Its History.* E. Hairer, G. Wanner. Undergrad. Texts in Math. Springer-Verlag, 1996, x + 374 pp, $42. [ISBN 0-387-94551-2] The traditional order of topics—sets, limits, continuity, derivatives, integrals—is reversed to reflect the historical development of the subject. Quotations and diagrams from original sources add to the historical flavor. Treats topics with the mathematical rigor of the period. This approach successfully

motivates a rigorous treatment of convergence, but provides difficult challenges to an instructor organizing a course. SA

Algebraic Topology, P*. *Homotopy Theory and Models.* Marc Aubry. DMV Seminar, Band 24. Birkhäuser Boston, 1995, ix + 117 pp, $33 (P). [ISBN 0-8176-5185-3] Condensed handbook provides an overview of homotopy theory from the point of view of algebraic models, especially the Sullivan and Quillen models of rational homotopy. TR

Topology, T*(16: 2). *Knots and Surfaces.* N.D. Gilbert, T. Porter. Oxford Univ Pr, 1994, xi + 268 pp, $46. [ISBN 0-19-853397-7] Introduction to knot theory and its interaction with the theory of surfaces and group presentations. Very nice intuitive introduction to invariants, including the Jones polynomial, via diagrams. Includes introductory material on topological spaces, surfaces, and group presentations. TR

Topology, T(15: 1). *A General Topology Workbook.* Iain T. Adamson. Birkhäuser Boston, 1996, viii + 152 pp, $26.50 (P). [ISBN 0-8176-3844-X] Introduction to point-set topology. Part I consists of definitions, statements of theorems, and 218 exercises on basics of topological spaces, mappings, compactness, and connectedness. Part II contains solutions to the exercises. Ideal for independent study. TR

Elementary Statistics, T(13: 1). *Statistics and Data Analysis: An Introduction, Second Edition.* Andrew F. Siegel, Charles J. Morgan. Wiley, 1996, xiii + 635 pp, $67.95. [ISBN 0-471-57424-4] Exploratory data analysis and Minitab computer output incorporated throughout. This edition features more categorical data and regression analysis, and more exercises. (*First Edition*, TR, October 1988.) RS

Elementary Statistics, T*(13: 1). *Foundations of Statistics.* Warren Hawley. Saunders College, 1996, xvi + 638 pp, $48. [ISBN 0-03-098253-7] Standard topics, divided equally between descriptive statistics, data analysis and probability topics, and inferential statistics topics. Many exercises based on real data. Unique features include Technology Tools sections with instructions for using Minitab and/or a graphing calculator (TI–82), references to appropriate programs in the video series *Against All Odds: Inside Statistics*, group exercises, and case studies involving writing and library work. RSK

Elementary Statistics, T(13–14: 1), S*. *Workshop Statistics.* Allan J. Rossman. Springer-Verlag, 1996, xx + 452 pp, $26.95 (P). [ISBN 0-387-94497-4] Interactive, activity-based approach to introductory statistics. Covers descriptive statistics, sampling, simple linear regression, hypothesis-testing for means and proportions (parametric tests only), experimental design. Activities use data collected by students or from popular media sources. Assumes access to statistical software (preferred) or graphing calculator (acceptable). LB

Elementary Statistics, T(13: 1). *Understanding Statistics, Fourth Edition.* Arnold Naiman, Robert Rosenfeld, Gene Zirkel. McGraw-Hill, 1996, xxii + 548 pp, (P). [ISBN 0-07-045915-0] Fun to read, but deceptively thorough in its treatment of elementary statistics. Delightful, full of thought- and laugh-provoking examples. Nice coverage of probability and nonparametrics. I particularly liked the initial class survey and follow-up throughout the book, and the "field projects" for students. (*Second Edition*, TR, November 1977.) KS

Elementary Statistics, T*(13–14: 1), C. *Statistics: A Bayesian Perspective.* Donald A. Berry. Duxbury Pr, 1996, x + 518 pp, $56.75, [ISBN 0-534-23472-0]; *Student Solutions Manual*, 105 pp, (P). [ISBN 0-534-23476-3] Possibly the first elementary text (assumes only high school algebra) to use a Bayesian approach! Appendices to most chapters give Minitab commands, including those needed to execute special Bayesian macros included on disk. Filled with examples and exercises drawn from medicine, science, and sports. RSK

Mathematical Statistics, T(15–16: 2). *Wahrscheinlichkeitsrechnung und mathematische Statistik.* Otfried Beyer, *et al.* Mathematik für Ingenieure und Naturwissenschaftler. BG Teubner Leipzig, 1995, 264 pp, DM 26,80 (P). [ISBN 3-8154-2075-X] Sophisticated introduction to probability, statistics. Few proofs, many examples with discussions, some exercises with solutions. JD-B

Mathematical Statistics, T(18: 1), P, L. *Multilevel Statistical Models, Second Edition.* Harvey Goldstein. Kendall's Lib. of Stat., V. 3. Edward Arnold (Copub: Halsted Pr), 1995, xiv + 178 pp, $49.95. [ISBN 0-340-59529-9] *First Edition* published in 1987 as *Multilevel Models in Educational and Social Research.* Concerned with "techniques for the analysis of highly structured data, both hierarchies and cross classifications." No exercises. RSK

Mathematical Statistics, T*(15–16: 2). *Mathematical Statistics with Applications, Fifth Edition.* Dennis D. Wackerly, William Mendenhall III, Richard L. Scheaffer. Duxbury Pr, 1996, xvii + 798 pp, $75.50. [ISBN 0-534-20916-5] Modest revision of the 1990 *Fourth Edition.* Major change is the expansion and updating of

exercises based on real data. A solid classical text. RSK

Statistical Methods, P., *Bayesian Analysis in Statistics and Econometrics.* Eds: Donald A. Berry, Kathryn M. Chaloner, John K. Geweke. Ser. in Prob. & Stat. Wiley, 1996, xxii + 577 pp, $79.95. [ISBN 0-471-11856-7] 48 papers on Bayesian analysis and its applications. Sections on forecasting and probability assessment; inference, estimation, and prediction; regression, linear models, and multivariate analysis; model selection; computation; applications; reliability and clinical trials; philosophical issues. RS

Statistical Methods, T(17–18), C, P. *Introduction to Graphical Modelling.* David Edwards. Texts in Stat. Springer-Verlag, 1995, xii + 274 pp, $39, with disk. [ISBN 0-387-94483-4] Application-oriented introduction with informal theoretical coverage using the PC-based MIM computer program. Emphasizes conditional independence model structure. Methods for discrete, continuous, and mixed models, hypothesis testing for mixed models, model selection and assessment, and applications of the EM-algorithm. Includes a MIM reference guide, and a student version of MIM on disk. RS

Statistical Methods, T(16–17). *Time Series Analysis: Forecasting and Control, Third Edition.* George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel. Prentice Hall, 1994, xvi + 598 pp. [ISBN 0-13-060774-6] *Third Edition* of classic text (1976 Holden Day edition, TR, January 1977). New edition includes rewrite of ARMA estimation and new sections on model specification, model selection criteria, tests for nonstationarity, state-space representations of ARMA models, structural, deterministic components in time series models, and intervention and outlier analysis. Numerous exercises and data sets (not on disk). MK

Computer Systems, C, P. *Publish It on the Web! Macintosh Version.* Bryan Pfaffenberger. Academic Pr, 1996, xxiii + 436 pp, $34.95 (P), with CD–ROM, [ISBN 0-12-553144-3]; *Publish It on the Web! Windows Version,* 1996, xxii + 441 pp, $34.95 (P), with CD–ROM. [ISBN 0-12-553140-0] Tips, techniques, and software tools for making documents and multimedia material available on the World Wide Web.

Applications (Economics), T(13: 3–4). *Mathematics: An Applied Approach, Sixth Edition.* Abe Mizrahi, Michael Sullivan. Wiley, 1996, xix + 1059 pp, $75.95. [ISBN 0-471-10701-8] Major topics: linear algebra, probability, calculus. Clear development. Many business applications. Takes advantage of graphing calculators. Many problem sets include examples from actuarial and CPA exams. (*Fourth Edition,* TR, February 1989.) KS

Applications (Economics), T(17–18: 1). *Applied Econometric Time Series.* Walter Enders. Wiley, 1995, xi + 433 pp, $55.95. [ISBN 0-471-03941-1] A lovely, thorough, concise treatment of time series methods. The text is approachable, the organization unique and useful, the examples insightful, and the exercises plentiful. The introduction to difference equations is a welcome addition to a time-series text. No data sets. MK

Applications (Engineering), S(15). *Engineering Mathematics with Maple.* John S. Robertson. Intern. Ser. in Pure & Appl. Math. McGraw-Hill, 1996, xiv + 278 pp, (P). [ISBN 0-07-053120-X] 27 labs cover a broad range of mathematical topics including Laplace transformations, Fourier series, and conformal mappings. Each lab contains applications and exploratory exercises. TR

Applications (Engineering), T(16–17: 1). *Reliability: Probabilistic Models and Statistical Methods.* Lawrence M. Leemis. Prentice Hall, 1995, xv + 319 pp. [ISBN 0-13-720517-1] Explains "how the mathematical models and results apply to engineering design and the analysis of lifetime data sets." Emphasizes reliability engineering, but models and methods described also apply to actuarial science and biostatistics. RSK

Applications, P. *The Nature of Statistical Learning Theory.* Vladimir N. Vapnik. Springer-Verlag, 1995, xv + 188 pp, $39.95. [ISBN 0-387-94559-8] Technical treatise concentrating on "the main results of learning theory and their connections to fundamental problems in statistics." Reviews the historical development of the mathematical analysis of learning processes. RSK

Applications, T(15–16: 1). *Life Insurance Mathematics, Second, Expanded Edition.* Hans U. Gerber. Springer-Verlag, 1995, xv + 217 pp, $59. [ISBN 3-540-58858-2] Expansion of the *First Edition* (TR, March 1991) by the addition of 75 pages of exercises and solutions by Samuel H. Cox. RSK

Reviewers

SA: Steve Abbott, St. Olaf; LB: Lynne Baur, Carleton; LC: Laura Chihara, St. Olaf; JD-B: John Dyer-Bennet, Carleton; TH: Tom Halverson, Macalester; MK: Michael Kahn, St. Olaf; RSK: Richard S. Kleber, St. Olaf; RM: Richard Molnar, Macalester; TR: Tommy Ratliff, St. Olaf; RS: Richard Single, St. Olaf; KS: Karla Spriestersbach, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; MW: Martha Wallace, St. Olaf.

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 103, Number 8 October 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

Return to the Riemann Integral

Robert G. Bartle

To J. T. Schwartz, on his 65th birthday

§1. INTRODUCTION. It is well known that the Riemann integral is not adequate for advanced mathematics, since there are many functions that are not Riemann-integrable, and since the integral does not possess sufficiently strong convergence theorems. To correct these deficiencies, Lebesgue developed his integral around the turn of the present century, and his integral has become the "official" integral in mathematical research.

However, there are also difficulties with the Lebesgue integral:

(1) There exist functions F that are differentiable at every point, but such that their derivatives F' are not Lebesgue integrable. Thus an added hypothesis is necessary to validate the formula

$$\int_a^b F' = F(b) - F(a). \tag{1a}$$

As one consequence, theorems justifying the substitution formula

$$\int_{\varphi(a)}^{\varphi(b)} f = \int_a^b (f \circ \varphi)\varphi' \tag{1b}$$

become unnecessarily complicated.

(2) Some improper integrals, such as the important Dirichlet integral

$$\int_0^\infty \frac{\sin x}{x}\,dx, \tag{1c}$$

do not exist as Lebesgue integrals (since $|x^{-1} \sin x|$ is not Lebesgue integrable).

(3) A considerable amount of measure theory needs to be developed before the Lebesgue integral can be defined.

It is the position of the present author that *the time has come to discard the Lebesgue integral as the primary integral.* We should replace it with a general form of the Riemann integral that—surprisingly enough—is *more* general than the Lebesgue integral and corrects the above difficulties. This generalization was discovered by Jaroslav Kurzweil and Ralph Henstock around 1960, but for some reason it has not become well known. Its definition is "Riemann-like", but its power is "super-Lebesgue". It is our view that we should not try to teach proofs to beginning calculus students, but that we should equip them with theorems to apply. Somewhat later, serious undergraduate students should be expected to understand appropriate proofs. We believe that most American undergraduates are not ready to study the Lebesgue integral, but that they are capable of

mastering a (somewhat stripped-down) version of the generalized Riemann integral. In §§2–10, we will provide an outline of such a version, with some side remarks to those who already know about the Lebesgue integral.

Historical remark. Over 20 years ago, E. J. McShane **[7]** made an eloquent argument for replacing the usual measure-theoretic approach to the Lebesgue integral by a Riemann-type approach that is afforded by the generalized Riemann integral. He later published a book **[8]** that could be used as a text for undergraduates in such a course. It is the present author's opinion that McShane was (i) overly optimistic in believing that the full Lebesgue integral can be taught to undergraduates, and (ii) overly conservative in developing the Lebesgue integral and not the generalized Riemann integral, which is more powerful and, we believe, conceptually simpler.

§2. BASIC DEFINITIONS. For the sake of simplicity, we will limit most of our remarks to the case of an interval $I := [a, b]$, $a < b$, in $\boldsymbol{R}$ and functions with (finite!) values in $\boldsymbol{R}$.

A *partition* of I is a finite collection of non-overlapping nondegenerate closed intervals $\{I_i\}_{i=1}^n$ whose union is I. Usually the partition is ordered and the intervals are specified by their end points; thus $I_i := [x_{i-1}, x_i]$, where

$$a = x_0 < x_1 < \cdots < x_{i-1} < x_i < \cdots < x_n = b. \tag{2a}$$

A *tagged partition* $P := \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is a finite set of ordered pairs, where the closed intervals $I_i = [x_{i-1}, x_i]$ form a partition of I and the numbers $t_i \in I_i$ are called the corresponding *tags*. If $P := \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is a tagged partition of I and $f\colon I \to \boldsymbol{R}$ is a function, then the *Riemann sum* $S(f; P)$ of f corresponding to P is the number

$$S(f;P) := \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}). \tag{2b}$$

The usual definition of the Riemann integral can be phrased: The number $A \in \boldsymbol{R}$ is the *Riemann integral* of $f\colon I \to \boldsymbol{R}$ if for every $\varepsilon > 0$ there exists a *constant* $\delta_\varepsilon > 0$ such that if $P := \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is any tagged partition of I satisfying $0 < x_i - x_{i-1} \le \delta_\varepsilon$ for $i = 1, \ldots, n$, then

$$|S(f;P) - A| \le \varepsilon. \tag{2c}$$

It turns out that the use of a *constant* $\delta_\varepsilon > 0$ restricts the Riemann integral quite considerably. The generalized Riemann integral is obtained by allowing δ_ε to be any strictly positive *function* on I. At first glance, that change seems to be very minor, but it turns out to make a profound difference in the properties of the resulting integral.

A strictly positive function δ on I is called a *gauge* on I. If δ is a gauge on I and $P := \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is a tagged partition of I, we say that P is *δ-fine* in case

$$0 < x_i - x_{i-1} \le \delta(t_i) \qquad \text{for} \quad i = 1, \ldots, n. \tag{2d}$$

The Nested Intervals Theorem implies that, given any gauge δ on I, there exist δ-fine partitions of I. The definition of the generalized Riemann integral differs from that of the ordinary Riemann integral by allowing nonconstant gauges.

(2.1) Definition. A number $B \in \boldsymbol{R}$ is the *generalized Riemann integral* of a function $f\colon I \to \boldsymbol{R}$ if for every $\varepsilon > 0$ there exists a gauge δ_ε on I such that if $P := \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is any partition of I that is δ_ε-fine, then

$$|S(f;P) - B| \le \varepsilon. \tag{2e}$$

In this case we will write $f \in \mathscr{R}^*(I)$ and denote $B = \int_I f = \int_a^b f$.

To show directly that $f \in \mathscr{R}^*(I)$, one must produce a suitable gauge δ_ε on I for any given $\varepsilon > 0$. However, there is a Cauchy condition for integrability and it is usually more convenient to use that condition (or other theorems) to establish the integrability of functions. It is an easy exercise to show that if $f \in \mathscr{R}^*(I)$, then the number B in (2e) is uniquely determined. Further, one can change the values of an integrable function on a *null* set without affecting the integrability or the value of the integral. The collection $\mathscr{R}^*(I)$ is a vector space and admits pointwise multiplication by functions of bounded variation. (Recall the Abel and Dirichlet Tests for non-absolutely convergent series.)

Remark. In establishing the details of the theory, it is found to be convenient to use a slightly different definition of δ-fineness.

§3. SOME EXAMPLES. We now give some examples of functions that belong to the collection $\mathscr{R}^*(I)$.

(3.1) Every Riemann integrable function on I is in $\mathscr{R}^*(I)$.

This follows from the fact that the gauge can be a strictly positive constant function. Thus, every continuous function on I is in $\mathscr{R}^*(I)$, and every step function on I is in $\mathscr{R}^*(I)$.

(3.2) If $h\colon [0,1] \to \boldsymbol{R}$ is Dirichlet's function (= the characteristic function of the rational numbers in $[0,1]$), then $h \in \mathscr{R}^*([0,1])$ and $\int_0^1 h = 0$.

To prove this assertion, we will define an appropriate gauge δ_ε. First we enumerate these rational numbers as $\{r_1, r_2, \dots\}$. We define $\delta_\varepsilon(r_i) := \varepsilon/2^{i+1}$, and if $x \in [0,1]$ is irrational we define $\delta_\varepsilon(x) := 1$; clearly δ_ε is a gauge on $[0,1]$. If P is a δ_ε-fine tagged partition, there can be at most two subintervals in P that have the number r_i as tag, and the length of each of those subintervals is $\le \varepsilon/2^{i+1}$. Hence the contribution to $S(h;P)$ from subintervals with tag r_i is $\le \varepsilon/2^i$. Since the terms in $S(h;P)$ with tags at irrational points contribute 0, we readily see that

$$0 \le S(h;P) < \sum_{i=1}^{\infty} \varepsilon/2^i = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows that $h \in \mathscr{R}^*([0,1])$ and that $\int_0^1 h = 0$.

(3.3) Every Lebesgue integrable function on I is in $\mathscr{R}^*(I)$.

A proof of this (non-obvious) result requires an understanding of the Lebesgue integral. Thus the teacher will want to know a proof, but the student is not concerned with this result or its proof.

(3.4) There exist functions in $\mathscr{R}^*(I)$ that do not belong to $\mathscr{L}(I)$.

Indeed, the function $F(x) := x^2\cos(\pi/x^2)$ for $x \in (0,1]$ and $F(0) := 0$ is readily seen to be differentiable at every point of $[0,1]$. It will be seen in §4 that this implies that $f := F' \in \mathscr{R}^*([0,1])$. [However, since F is not absolutely continuous on $[0,1]$ the teacher will understand that $f \notin \mathscr{L}([0,1])$.]

(3.5) If $\Sigma_{n=1}^{\infty} a_n$ is any convergent series, then one can define $h(x) := 2^n a_n$ for $x \in (1/2^n, 1/2^{n-1}]$ for $n \in N$ and $h(0) := 0$. A gauge can be constructed to show that $h \in \mathscr{R}^*([0,1])$ and that

$$\int_0^1 h = \sum_{n=1}^{\infty} a_n.$$

Moreover, $|h| \in \mathscr{R}^*([0,1])$ if and only if the series $\Sigma_{n=1}^{\infty} a_n$ is absolutely convergent [if and only if $h \in \mathscr{L}([0,1])$].

Example (3.5) shows that the absolute value of function in $\mathscr{R}^*(I)$ is not necessarily in $\mathscr{R}^*(I)$. Thus the generalized Riemann integral is not an "absolute integral". That is why the Dirichlet integrand in (1c) can be in $\mathscr{R}^*([0,\infty))$.

§4. THE FUNDAMENTAL THEOREM. We have noted in §1 that the Lebesgue integral is not powerful enough to integrate *every* derivative. That fact led Denjoy and Perron to develop their (very different) theories of integration. The details and subtleties of these theories of integration are quite considerable (see [2]). This stands in marked contrast with the generalized Riemann integrals, for which, as we will now see, the details are remarkably simple.

(4.1) Fundamental Theorem. *If F: $[a,b] \to \mathbf{R}$ is differentiable at every point of $I := [a,b]$, then $f = F'$ belongs to $\mathscr{R}^*(I)$ and*

$$\int_a^b f = F(b) - F(a). \tag{4a}$$

Proof: If $t \in I$, since $f(t) = F'(t)$ exists, given $\varepsilon > 0$ there exists $\delta_\varepsilon(t) > 0$ such that if $0 < |z - t| \leq \delta_\varepsilon(t)$, $z \in I$, then

$$\left| \frac{F(z) - F(t)}{z - t} - f(t) \right| \leq \varepsilon.$$

Thus a gauge δ_ε has been defined on I. Further, if $|z - t| \leq \delta_\varepsilon(t)$, $z \in I$, then

$$|F(z) - F(t) - (z - t)f(t)| \leq \varepsilon|z - t|.$$

Hence, if $a \leq u \leq t \leq v \leq b$ and $0 < v - u \leq \delta_\varepsilon(t)$, then it follows that

$$\begin{aligned}
&|F(v) - F(u) - (v - u)f(t)| \\
&\quad \leq |F(v) - F(t) - (v - t)f(t)| + |F(t) - F(u) - (t - u)f(t)| \\
&\quad \leq \varepsilon(v - t) + \varepsilon(t - u) = \varepsilon(v - u).
\end{aligned}$$

If $P = \{([x_{i-1}, x_i], t_i)\}_{i=1}^n$ is a δ_ε-fine partition of I, then the telescoping sum $F(b) - F(a) = \Sigma_{i=1}^n \{F(x_i) - F(x_{i-1})\}$ satisfies the approximation

$$\begin{aligned}
|F(b) - F(a) - S(f;P)| &= \left| \sum_{i=1}^n \{F(x_i) - F(x_{i-1}) - f(t_i)(x_i - x_{i-1})\} \right| \\
&\leq \sum_{i=1}^n |F(x_i) - F(x_{i-1}) - f(t_i)(x_i - x_{i-1})| \\
&\leq \sum_{i=1}^n \varepsilon(x_i - x_{i-1}) = \varepsilon(b - a).
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this shows that f is in $\mathscr{R}^*(I)$ and establishes (4a).

It is not difficult to extend the Fundamental Theorem (4.1) to a function that is the derivative of a continuous function at all but a countable set of points in I. Thus, it follows that the function defined by $f(x) := 1/\sqrt{x}$ for $x \in (0,1]$ and $f(0) := 0$ is in $\mathscr{R}^*([0,1])$, since it is the derivative of the function $F(x) := 2\sqrt{x}$ except at $x = 0$. Thus we have

$$\int_0^1 f = \int_0^1 \frac{1}{\sqrt{x}}\,dx = 2\sqrt{x}\,\Big|_0^1 = 2.$$

Of course, the function f is well known to have an *improper* Riemann integral; we have just seen that it has an ordinary *generalized* Riemann integral.

§5. SUBSTITUTION THEOREMS. In view of the simplicity of the Fundamental Theorem (4.1), one can make corresponding improvements in the theorems justifying the familiar "substitution formula":

$$\int_{\varphi(a)}^{\varphi(b)} f = \int_a^b (f \circ \varphi)\,\varphi'. \tag{5a}$$

Although this formula is a basic tool in analysis, it is seldom stated (or proved) with the generality required for nontrivial use. We will content ourself with two theorems that the generalized Riemann integral renders valid.

(5.1) Substitution Theorem, I. *Let φ: $[a,b] \to \boldsymbol{R}$ be differentiable on $I := [a,b]$ and let F be differentiable on the interval $\varphi(I)$. If $f(x) = F'(x)$ for all $x \in \varphi(I)$, then equation* (5a) *holds.*

Proof: It follows from the Chain Rule that $(F \circ \varphi)'(x) = (f \circ \varphi)(x)\varphi'(x)$ for all $x \in I$. Two applications of the Fundamental Theorem (4.1) imply that

$$\int_a^b (f \circ \varphi)\,\varphi' = F \circ \varphi\big|_a^b = F\big|_{\varphi(a)}^{\varphi(b)} = \int_{\varphi(a)}^{\varphi(b)} f.$$

The proof of the next result is more subtle.

(5.2) Substitution Theorem, II. *Let φ be a strictly increasing and differentiable mapping of $I := [a,b]$ onto $\varphi(I) = [\varphi(a), \varphi(b)]$. Then f belongs to $\mathscr{R}^*(\varphi(I))$ if and only if $(f \circ \varphi)\varphi'$ belongs to $\mathscr{R}^*(I)$. In this case* (5a) *holds.*

Both (5.1) and (5.2) can be extended to more general circumstances.

§6. IMPROPER INTEGRALS. One of the remarkable properties of the generalized Riemann integral (that is not shared by either the ordinary Riemann integral or the Lebesgue integral) is the following theorem due to H. Hake.

(6.1) Hake's Theorem. *A function f belongs to $\mathscr{R}^*([a,b])$ if and only if it belongs to $\mathscr{R}^*([a,c])$ for every $c \in (a,b)$ and $\lim_{c \to b-} \int_a^c f$ exists in $\boldsymbol{R}$. In this case $\int_a^b f = \lim_{c \to b-} \int_a^c f$.*

One can interpret Hake's Theorem as asserting: *The generalized Riemann integral cannot be extended by adjoining functions with "improper integrals"*. In other words, if the "improper integral" exists, then the integral exists as (an ordinary) generalized Riemann integral.

The student would be interested in the half of the theorem asserting that the integral can be evaluated as a limit; the proof of that part is rather easy. The harder part of the proof is of interest only to the teacher, since only the teacher believes in improper integrals.

§7. CHARACTERIZATION OF INDEFINITE INTEGRALS. In §4 we discussed one aspect of the Fundamental Theorem, namely the integrability of any derivative. The other aspect of the Fundamental Theorem pertains to the differentiation of the *indefinite integral of f*, which is the function F defined by

$$F(x) := \int_a^x f \qquad \text{for} \quad x \in [a, b]. \tag{7a}$$

In an undergraduate course, it would probably be best to content oneself with showing that $F'(c) = f(c)$ at every point $c \in I$ where f is continuous.

[The teacher, of course, should know more. In fact, using the Vitali Covering Theorem, one can show that F is differentiable almost everywhere and that

$$F'(x) = f(x) \quad \text{almost everywhere.} \tag{7b}$$

However, the proof of this fact is a bit too much for most undergraduates. The teacher should know that there is an extension of Lebesgue's characterization of indefinite integrals that is valid for the generalized Riemann integral. Somewhat imprecisely stated, a function F is an indefinite integral of $f \in \mathscr{R}^*(I)$ if and only if (i) $F'(x) = f(x)$ almost everywhere in I, and (ii) on the set where (i) does not hold, then F has "arbitrarily small variation". This characterization can be used to give a proof that $\mathscr{L}(I) \subset \mathscr{R}^*(I)$, of interest to the teacher. It also implies that $f \in \mathscr{L}(I)$ if and only if both f and $|f|$ belong to $\mathscr{R}^*(I)$.]

However, from the student's standpoint, it would be appropriate merely *to define* the space of *Lebesgue integrable functions* to be:

$$\mathscr{L}(I) := \{f \in \mathscr{R}^*(I) : |f| \in \mathscr{R}^*(I)\},$$

and to make $\mathscr{L}(I)$ into a semi-normed space under

$$\|f\|_1 := \int_I |f|.$$

§8. CONVERGENCE THEOREMS. One of the main reasons for the interest in the Lebesgue integral is its convergence theorems. It is quite surprising that they also hold in $\mathscr{R}^*(I)$.

It is easy to prove that if (f_n) is a sequence in $\mathscr{R}^*([a, b])$ that converges *uniformly* to f on $[a, b]$, then $f \in \mathscr{R}^*([a, b])$ and

$$\int_a^b f = \lim_{n \to \infty} \int_a^b f_n. \tag{8a}$$

However, the Monotone Convergence Theorem is also true in $\mathscr{R}^*(I)$.

(8.1) Monotone Convergence Theorem. *Let (f_n) be a sequence in $\mathscr{R}^*([a, b])$ that is monotone increasing*:

$$f_1(x) \le \cdots \le f_n(x) \le f_{n+1}(x) \le \cdots \qquad \textit{for all} \quad x \in [a, b],$$

and let $f(x) = \lim_n f_n(x) \in \boldsymbol{R}$ for all $x \in [a, b]$. Then $f \in \mathscr{R}^*([a, b])$ if and only if

$$\sup_n \int_a^b f_n < \infty.$$

In this case, (8a) *holds.*

From this one can prove Fatou's Lemma and the following version of the Dominated Convergence Theorem.

(8.2) Dominated Convergence Theorem. *Let (f_n) be a sequence in $\mathscr{R}^*([a,b])$, let $g, h \in \mathscr{R}^*([a,b])$ be such that*

$$g(x) \leq f_n(x) \leq h(x) \qquad \textit{for all} \qquad x \in [a,b],$$

and let $f(x) = \lim_n f_n(x) \in \boldsymbol{R}$ for all $x \in [a,b]$. Then $f \in \mathscr{R}^*([a,b])$ and (8a) *holds.*

These proofs do not need any measure theory, but they may be slightly out of the range of most undergraduates.

§9. MEASURE THEORY. Ultimately, it is desirable that students learn some measure theory. We now suggest how that theory can be developed from the generalized Riemann integral. (The situation is slightly more complicated for an infinite interval.)

As usual, we define a *null set* in $I := [a,b]$ to be a set that can be covered by a countable union of intervals with arbitrarily small total length. We define a function $f\colon I \to \boldsymbol{R}$ to be *measurable* if there exists a sequence of step (or continuous) functions on I that converges to f almost everywhere (that is, on the complement of a null set). One can now relate this notion to the generalized Riemann integral.

(9.1) Measurability Theorem. *Every $f \in \mathscr{R}^*(I)$ is measurable on I.*

Indeed, f is equal almost everywhere to the limit of a sequence of difference quotients of its (continuous) indefinite integral function.

(9.2) Integrability Theorem. *If $f\colon I \to \boldsymbol{R}$ is measurable on I and if there exist $g, h \in \mathscr{R}^*(I)$ such that $g(x) \leq f(x) \leq h(x)$ for all $x \in I$, then $f \in \mathscr{R}^*(I)$.*

The proof uses the fact that f is the limit almost everywhere of a sequence of simple functions and the Dominated Convergence Theorem (8.2).

We say that a set $A \subseteq I = [a,b]$ is *measurable* if its characteristic function is a measurable function (or, equivalently, belongs to $\mathscr{R}^*(I)$). One can show that the sets

$$A \cap B, \quad A \cup B, \quad \text{and} \quad I - A$$

are measurable sets in I whenever A, B are measurable. Thus the collection $\mathscr{M}(I)$ of all measurable subsets of I is an algebra of sets, and the Monotone Convergence Theorem (8.1) implies that $\mathscr{M}(I)$ is a σ-algebra of sets. Since $\mathscr{M}(I)$ contains all intervals in I, it follows that it contains the *Borel measurable* subsets of I. Since $\mathscr{M}(I)$ contains all null subsets of I, it follows that it is precisely the collection of all *Lebesgue measurable* subsets of I.

§10. FINAL COMMENTS. It is easy to see that everything extends to complex-valued functions, or to functions with values in $\boldsymbol{R}^m$, $m > 1$.

The theory can also be extended to functions whose domain is a non-compact interval by using a simple device that is discussed in the books of McLeod and of DePree and Swartz.

The main outlines of the theory carry over easily for functions defined on a compact rectangle in $\boldsymbol{R}^m$, $m > 1$; see the books of Mawhin, McLeod, and Pfeffer that are cited below. There are certain complications when the domain is not compact, but the major parts of the theory extend. One of the active areas of research in this topic is in adapting the integral so that a version of the Divergence Theorem with minimal hypotheses holds. The reader is referred to the book of Pfeffer for an account of this work, and to papers of Jarník, Jurkat, Kurzweil, Mawhin, Nonnenmacher, Pfeffer and others for more detail. Some very significant results have been obtained in this direction, but it seems fair to say that a completely satisfactory theory has not yet been established.

In the preceding discussion the domains of the functions have been assumed to belong to one of the spaces $\boldsymbol{R}^m$. Some important steps have been taken to extend the theory to more general domains; we refer the reader to the recent book of Henstock for more details and a very comprehensive bibliography.

There is an account of the history of this material in the books of McLeod and Henstock. The most complete account of the theory in $\mathscr{R}^*([a,b])$ is in the excellent recent book by Gordon, where it is proved that the generalized Riemann integral (there called the Henstock integral) is equivalent to the integrals of Denjoy and Perron.

REFERENCES

1. DePree, J. D. and Swartz, C. W., *Introduction to real analysis*, John Wiley & Sons, New York, 1988.
2. Gordon, R. A., *The integrals of Lebesgue, Denjoy, Perron, and Henstock*, Grad. Studies in Math., vol. 4, Amer. Math. Soc., Providence, 1994.
3. Henstock, R., *The general theory of integration*, Oxford Math. Monographs, Clarendon Press, Oxford, 1991.
4. Lee, Peng Yee, *Lanzhou lectures on Henstock integration*, World Scientific Pub. Co., Singapore, 1989.
5. Mawhin, J., *Analyse—fondaments, techniques, évolution*, DeBoeck Université, Brussels, 1992.
6. McLeod, R. M., *The generalized Riemann integral*, Carus Math. Monographs, no. 20, Math. Assn. Amer., Washington, 1980.
7. McShane, E. J., *A unified theory of integration*, Amer. Math. Monthly, **80** (1973), 349–359.
8. McShane, E. J., *Unified integration*, Academic Press, New York, 1983.
9. Pfeffer, W. F., *The Riemann approach to integration*, Cambridge Univ. Press, London, 1993.

Department of Mathematics
Eastern Michigan University
Ypsilanti, MI 48197
mth_bartle@emuvax.emich.edu

The History of the Hand-Held Electronic Calculator

Kathy B. Hamrick

The history of the electronic hand-held calculator is a missing chapter in the history of mathematics and computer science. This history is as obscure as the calculators are prevalent; yet it is an important aspect of our heritage as mathematicians. This article describes an interesting part of this calculator history; namely an insider's story of the invention of the first electronic hand-held calculator, called the CAL-TECH. The CAL-TECH is the prototype of the first commercial version of an electronic hand-held calculator, the Pocketronic. One CAL-TECH is now in the Smithsonian and another is in the Texas Instruments archives. Much of the information for this story was obtained from interviews with two of CAL-TECH's inventors.

To add further insight and inject humor into this history, this article also includes interesting and humorous sexist comments and facts about the early calculators. This information was found in newspaper and magazine articles from the years of the calculator's early history. The complete history of the electronic hand-held calculator begins with this story, continues through the development and shakeout of the resulting calculator industry, and still continues with today's technology.

THE INVENTION. In the fall of 1965, electronic calculators had to be plugged in (120 v), were the size of typewriters, and cost as much as an automobile. However, the first hand-held electronic calculator was not invented to replace these monsters. The invention was intended to introduce the world to the integrated circuit or microchip, make the term "chip" mean more to the average American than "potato chip", and thus increase Texas Instruments' market for integrated circuits.

The invention of the first hand-held electronic calculator was initiated in 1965 by Pat Haggerty, the president of Texas Instruments (TI). While working for Haggerty, one of TI's engineers, Jack Kilby, invented the integrated circuit, and TI then developed an early lead in the manufacture of microchips that contained integrated circuits. At that time, however, the market for chips was limited to the military's Minute Man missiles and to industrial markets. Even computer manufacturers were wary of the chip. TI's problem with the sale of the chips was compounded by their limited consumer market experience and the fact that the chips did not wear out. It sometimes took many attempts to make a "good" chip, but once a chip was good, it usually stayed good. If TI wanted to produce and sell more chips, they needed to find more products that used the chip. A dramatic invention was needed to launch the integrated circuit into everyday life.

Ten years earlier, Pat Haggerty had used an invention to successfully launch the transistor into everyday life. In 1954, he was Vice President of Texas Instruments

and TI led the world in the mass production of transistors. Computer manufacturers, television and radio manufacturers and the general public, however, still thought in terms of vacuum tubes and did not trust transistors. To remedy this problem, Haggerty ordered his engineers to develop a shirt pocket radio that used transistors. Haggerty's engineers invented the radio; but because of Texas Instruments' very limited consumer products experience, they joined with Regency Company of Indiana to market the radio. The first pocket radio was introduced just before Christmas, 1954, and over 100,000 radios were sold the first year. (Reid, 1984). This performance impressed big companies like IBM and these companies began to buy transistors from TI. The general public who bought the radios also began to trust transistor products and Texas Instruments thus became a world leader in the production and sale of transistors.

Haggerty decided to use the same invention technique to introduce the world to the integrated circuit. In the fall of 1965, Pat Haggerty and Jack Kilby, the inventor of the integrated circuit, were returning to Dallas on a plane. According to Kilby, Haggerty suggested he invent a calculator that would fit in a shirt pocket like the radio, or invent a lipstick-size dictaphone machine, or invent something else that used the microchip. Kilby decided the calculator would be more interesting. By the time the plane landed, the first electronic hand-held calculator had been suggested, or ordered, into existence as a step-child of the integrated circuit.

When Haggerty allowed Kilby to choose his own team of engineers, he chose a young self-taught engineer, Jerry Merryman, as the project manager. Merryman had attended Texas A & M for a few years, but had never graduated. However, Merryman was known as one of the brightest young engineers at TI. Another engineer, James Van Tassel, completed the team.

According to Merryman, Kilby first called several TI engineers together. He told them that he had been talking to Pat Haggerty on a plane, and Haggerty thought that the way electronics were going, we could have our own personal computer of sorts which would be portable, and would replace the slide rule. Merryman said that Haggerty's ideas were sometimes "off the wall", but since the ideas frequently paid off, they were taken seriously. Kilby told the engineers that the personal computer should be the size of a book which he had on his desk, it needed some buttons or some sort of input, it needed some neon lights or some sort of output, and it needed batteries for power. Not yet using the term "calculator", the group discussed building a *slide rule computer*. Van Tassel worked primarily with the keyboard, Kilby himself worked on the power source, and Merryman inherited the logic and the output.

A few weeks into the project, the TI accounting office demanded the project be given a name. By this time, the team was referring to the device as a calculator instead of a computer, so the project was code named, CAL-TECH. Kilby said that in retrospect, it was a dumb choice of a name. The project was supposed to be secret, and anyone at Texas Instruments could tell from the name that they were working on a calculator.

The project progressed with unusual speed. For desk top calculators, a large working model called a breadboard was developed first; the calculator was then patterned after the breadboard. The breadboard for the CAL-TECH, however, was developed simultaneously with the calculator; and the breadboard was used only to test chips. CAL-TECH's output device became the team's biggest problem. Neon tubes were fine for desk top calculators, but they would drain the CAL-TECH's batteries in a few minutes. TI was, at that time, working on Light Emitting Diodes (LEDs), but they weren't ready for commercial use. Merryman solved the

problem with a thermal printing device that burned numbers onto paper tape and used very little power. He continued this work and now holds a patent for a thermal heating device called an "Integrated Heater Element Array and Drive Matrix" (US Patent 3,501,615, March 17, 1970). Thus, unlike its predecessors, the CAL-TECH had a paper tape output instead of display lights.

About one year after the plane ride, in November, 1966, the CAL-TECH team showed a working calculator to Pat Haggerty. It was about 4" by 6" by 1.5" and was heavy (45 oz.) because it was made of a solid brick of aluminum. According to Merryman (1993), "We got a solid brick of aluminum and took a milling machine and hollowed it out." The output was a small paper tape that could be read through a window (if one had very good eyes), and the tape could be kept as a record. It had 18 keys: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ., ×, +, −, ÷, C, E, and P. It could add, subtract, multiply, and divide. Actually it really could only add and subtract. Multiplication was done by repeated addition and division was done by repeated subtraction. It was what is known in the industry as a "four-banger" or as "plain-vanilla", and it impressed Pat Haggerty a lot! He called in TI's lawyers and it took them longer to get the patent than it did for the CAL-TECH team to invent the calculator. The patent was first filed September 29, 1967, then refiled May 13, 1971, then refiled again on December 21, 1972. The patent number for the CAL-TECH is 3,819,921. If you examine the back of any TI calculator, the first patent number will be the patent number for CAL-TECH.

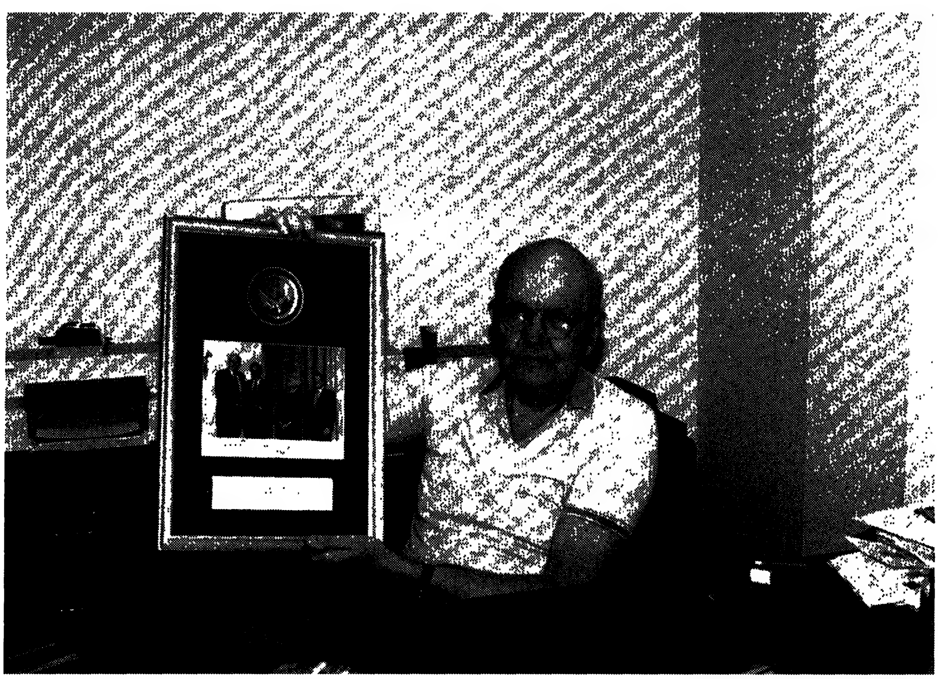


Figure 1. Dr. Jack Kilby with the National Medal of Technology, 1990. The award was given by President Bush.

Few inventors whose creations have changed the world as dramatically as the hand-held calculator are still alive and hence we know most of them only through books. Such is not the case with Jack Kilby, Jerry Merryman, and James Van Tassel. Jack Kilby is a private consultant in Dallas, Texas. Jerry Merryman is a

Figure 2. Mr. Jerry Merryman with the Pocketronic and the CAL-TECH. Interview, September, 1993.

Texas Instruments Fellow and also lives in Dallas. James Van Tassel is retired and lives in Ohio. All three have been very generous with their time for interviews.

THE POCKETRONIC. When the CAL-TECH was completed, Texas Instruments still had almost no experience with consumer products and was not sure it wanted any. Therefore, as it did with the pocket radio, TI joined with another company, Canon of Japan, to market the first calculator. This commercial version of the CAL-TECH was a much lighter calculator called the Pocketronic. Its case was plastic instead of metal, but the inside of the Pocketronic and the Cal Tech were essentially the same. Canon and Texas Instruments introduced the Pocketronic on April 14, 1970, the day before income tax returns were due.

Business Week magazine described it as "the portable, pocketable, all electronic consumer calculator that the electronics industry had long dreamed about". **[4]** The magazine also reported that three Canon engineers secretly spent three months in Dallas designing the Pocketronic. No mention, however, was made in the media of the calculator's inventors. The 1970 August edition of *Popular Science* magazine also described the Pocketronic. Both articles described the thermal tape output and the small size and weight—1.8 pounds. It was a large, expensive ($400) monster by today's standards, but in 1970, its only competition was very large desk top calculators that cost more than $2000. The Pocketronic's relatively modest $400 price tag was over $1500 in 1995 dollars, but a $2,000 desk top model would be over $7,500 in 1995 dollars.

It is hard to distinguish between today's calculators and computers. Graphing calculators have 32K memory, the equivalent of an early Apple computer. The

Figure 3. The "insides" of the Pocketronic and the CAL-TECH. Interview, September, 1993.

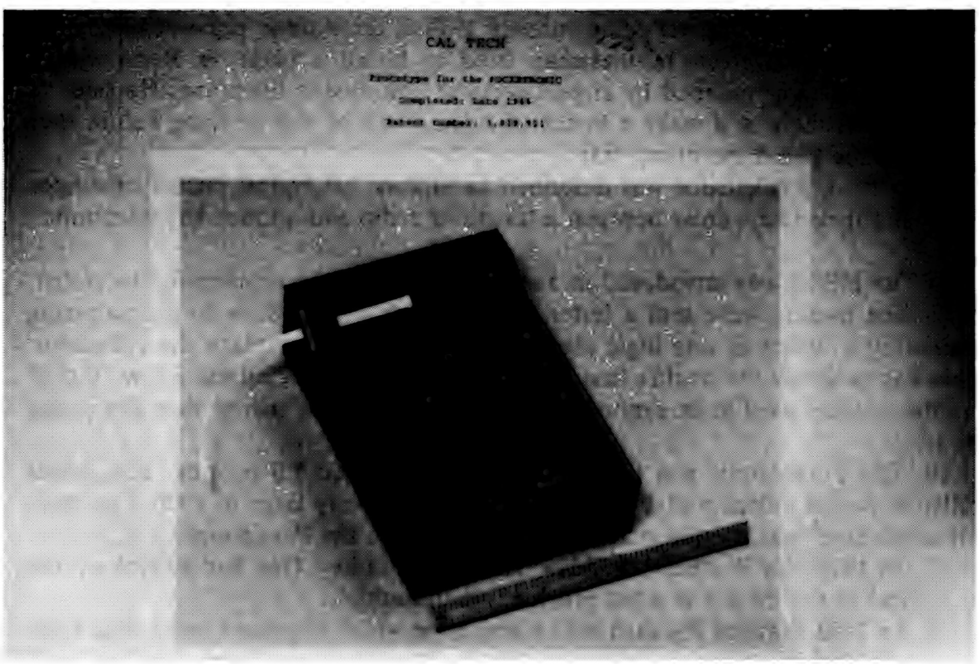


Figure 4. The CAL-TECH in the Smithsonian. Courtesy of Texas Instruments.

main distinguishing feature is the keyboard: a computer has a "qwerty" keyboard and a calculator has an alphabetic keyboard. Texas Instruments, however, has announced a new calculator with a "qwerty" keyboard. The distinction between the two will eventually disappear altogether. Today, a major difference is that the history and development of the computer is well documented, but the history and development of our powerful hand-held electronic calculators is still relatively unknown.

INTERESTING FACTS ABOUT EARLY CALCULATORS. The following information gathered from newspapers and magazines gives further insight and inject humor into the history of the hand-held calculator:

1. Edward Talko, manager of Keuffel & Esser, the oldest slide rule manufacturer, said in a 1977 interview, "The slide rule will still be around the way a horse is still around; but the calculator is taking its place the way the automobile replaced the horse." Mr. Talko used a calculator, but checked his answers with a slide rule [7].
2. Keuffel & Esser made its last slide rule in 1975.
3. The first solar powered calculator was the Teal Photon made by Teal Industries of California.
4. The HP-35 had only a degree mode, no radians.
5. In 1973, Consumer Reports gave a high rating to calculators with *click type keys* because one could hear a reassuring click when the buttons were pushed. Also, the LED displays caused glare problems, so calculators with a shaded display were rated high. ("Electronic Mini-Calculators", 1973).
6. HP introduced the HP-65 in January 1974 at the *modest* price of $795.
7. In 1972, a Detroit businessman tried to board a plane in Reno with a calculator. He was grabbed by airport security and missed his plane. He said, "I had to stand there and make a bunch of calculations on the — thing before they would let me board the plane." [3]
8. In 1973, a calculator was described as follows: "A typical pocket calculator looks and feels like a cross between a transistor radio and a touch-key telephone." [2]
9. The HP-35 was introduced in January, 1972 and was recalled in December, 1972. The owners were sent a letter pointing out idiosyncrasies in programming caused by a defect in one logic algorithm. HP offered to replace the calculator. This was probably the world's first instant recall. The defect caused a few 10 digit numbers, when used in an exponential function, to give an answer that was wrong by 1%.
10. The Pocketronic was 8″ × 4″ × 1.5″ and weighed 1.8 pounds. Yet, it was called a *pocket* calculator? Pockets must have been very large in 1970. The term "Bantam Size" was used in many articles to describe the Pocketronic.
11. In 1971, the Brother-310 ($349) had no divide key. One had to look up the reciprocal of the divisor in a list provided, then multiply.
12. In 1972, Hewlett Packard held a contest in which engineers using slide rules competed against engineers using the HP-35. The HP-35 group won, 65 sec. vs. 5 min. [8]
13. In 1973, the early HP-45s could be set in a timer mode. This was an accident of programming, and not all HP-35s had this "feature". Customers thought this was great, and sent back the ones that couldn't as defective. *A $395 stop watch*!

HUMOROUS SEXIST COMMENTS. In the 1970s, the press viewed the world of calculators as a man's domain as evidenced in the following comments:

1. "Calculators are being sold to engineers, college students, and women to use for shopping." [6]
2. "Every housewife will have one (calculator) when she goes shopping." [11]
3. "Salesmen use them to compute estimates and prices for carpeting and fences. A professional pilot carries one for navigational calculations. A housewife with skeet-shooting sons checks shooting record cards." [1]
4. "At the supermarket, the new calculator will help your wife find the best unit price bargains. At the lumberyard, they'll help you decide which combination of plywood, lumber and hardboard would be least expensive for your project." [2]
5. "Wives buy calculators as a gift for the man that has everything. But often, wives end up using them more than their husbands to balance the checkbook or keep track of household expenses." [13]
6. "Dad will use a pocket calculator in his work—whether he is an engineer, an insurance salesman, a garage manager or a business executive. Mom will have her own purse calculator, in one of several fashionable colors, which she will use in buying groceries, in balancing the checkbook, and in converting recipes." [5]

REFERENCES

1. Berger, I. (1973, February). Electronic calculators: how to choose the right one. *Popular Mechanics*, **139**, pp. 86–90.
2. Buckwalter, L. (1973, February). Now it's pocket calculators. *Mechanics Illustrated*, **69**, pp. 108–109.
3. Calculated boom. (1972, October). *Newsweek*, **80**, pp. 84–85.
4. A calculator to fit the pocket. (1970, April 18). *Business Week*, pp. 28–29.
5. Detman, A. (1992, September). Pocketing the profits. *Dun's*, **100**, pp. 89–90.
6. Eastman Kodak is increasing prices 6.5%. (1974, April 16). *New York Times*, p. 51.
7. Endangered species: The slide rule. (1977, March). *Science Digest*, **81**, pp. 68–69.
8. Frye, J. T. (1972, May). A versatile pocket calculator. *Popular Electronics*, **36**, pp. 58–61.
9. Kilby, Dr. Jack St. Clare. (1993, September). Interview with author.
10. Merryman, Mr. Jerry Dale. (1993, September 11). Interview with author.
11. Miller, R. (1973, September). Make a giant step forward with the latest in electronics technology —the mini-calculator. *Man–Science–Technology*, **33**, pp. 5–6.
12. Reid, T. R. (1984). *The Chip*. New York: Simon and Schuster.
13. Smith, W. D. (1972, August 20). Hand held calculators: tool or toy? *New York Times*, p. 7.

Department of Mathematics and Computer Science
Augusta College
Augusta, GA 30904-2200
khamrick@admin.ac.edu

The Road to Chaos is Filled with Polynomial Curves

Richard D. Neidinger and R. John Annen, III

1. INTRODUCTION. The bifurcation diagram (Figure 1) is much like an aerial photograph—there are many interesting things revealed but it's hard to see the detail or to interpret the features. By superimposing a family of polynomial curves (Figure 2), this picture turns into a road map—marking trails, boundaries, milestones, and labeling points of interest. Both the curves and the diagram are generated by iterating the function $x^2 + c$ as the parameter ranges across a horizontal c-axis from 0 to -2. The initial iterates form the curves with the nth curve being the nth iterate of the critical value zero as a function of c. The bifurcation diagram (also known as the orbit diagram, the Feigenbaum diagram, or the road to chaos) is a plot of later iterates, showing the eventual state of the system.

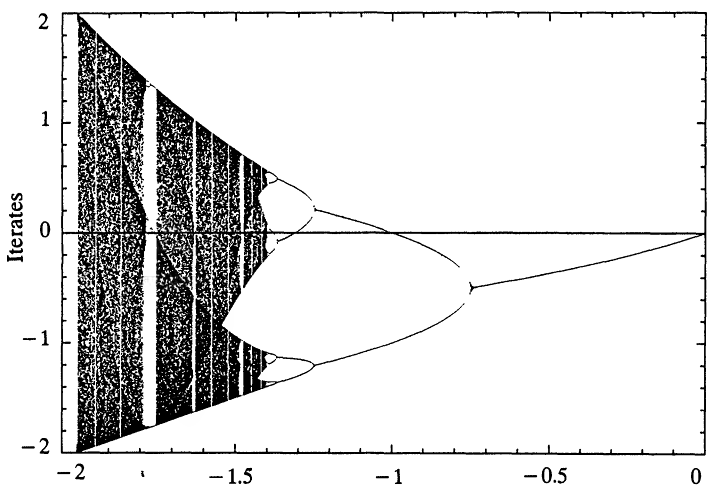


Figure 1. The bifurcation diagram of $f_c(x) = x^2 + c$.

In the first half of the paper, curve intersections provide insight into the dynamics of the system. The roots of the curves are shown to mark periodic windows in the bifurcation diagram, usually of invisible width, indicating a remarkable preponderance of attracting periodic phenomena that contrasts with the seeming chaos in the diagram. Other intersections mark parameter values where the iteration is chaotic. A brief detour into the complex plane yields a simple algorithm that color codes the Mandelbrot set to indicate the periods of the attracting phenomena. The second half of the paper studies how curve inequalities

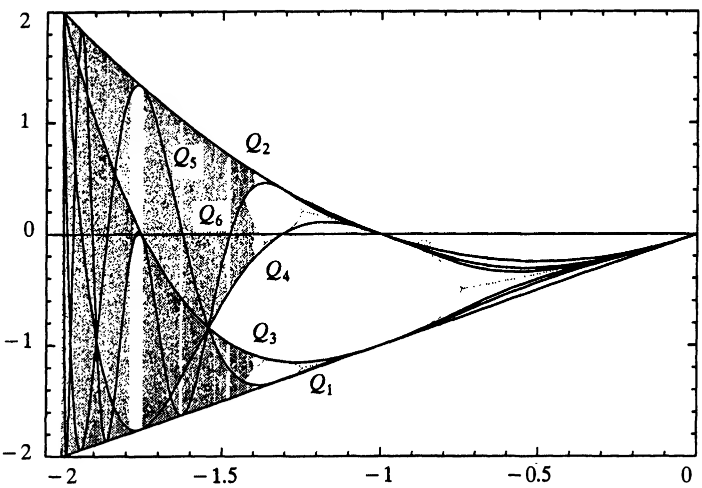


Figure 2. The first six Q-curves reveal dynamics and shapes within the diagram.

describe the shapes seen in the bifurcation diagram. We show why the high-numbered iterates condense along the polynomial curves. The curves also form envelopes that bound embedded, self-similar copies of the bifurcation diagram. Curve intersections specify the exact beginning and ending parameters for every envelope found in the diagram and, thus, mark where each window ends. Finally, Šarkovskiĭ's ordering can be seen in the order in which the curves first cross the axis. The ordering is argued without using Šarkovskiĭ's theorem.

Let's begin by defining terms and notation. For the standard dynamical system $f_c(x) = x^2 + c$, the iterates are defined by $x_{i+1} = f_c(x_i)$, or equivalently $x_i = f_c^i(x_0)$, for some fixed x_0. A sequence of such iterates is usually called an *orbit* and the *dynamics of the system* refers to the behavior of orbits. The *bifurcation diagram*, in Figure 1, displays the dynamics for different values along the horizontal c-axis by vertically plotting iterates, numbered 300 to 900, in the orbit of zero.

Define the nth polynomial $Q_n(c) = f_c^n(0)$ to be the nth iterate of f_c starting from zero. Specifically, $Q_1(c) = c$, $Q_2(c) = c^2 + c$, and $Q_3(c) = (c^2 + c)^2 + c$. In general, $Q_{n+1}(c) = (Q_n(c))^2 + c$. Thus, each $Q_n(c)$ is a 2^{n-1} degree polynomial. The graph of any Q_n as a function of c will be called a *Q-curve*. The first six Q-curves are displayed in Figure 2 and portions of these curves are visible as a relative density of points in Figure 1.

Q-curves are iterates of the critical value (in **[PJS]**, p. 633, they are called *critical-value lines*). The important role of the critical value is found in **Fatou's Theorem:** Every attracting cycle for a polynomial attracts at least one critical point **[Br]**. Even if the iteration for a bifurcation diagram starts from a different x_0, the eventual state reveals the same Q-curves. Indeed, the bifurcation diagram in the background of Figure 2 uses $x_0 = 0.2$, while the curves are iterates of the critical value zero. Typically, even bifurcation diagrams start with zero in order to avoid isolated values where x_0 may be a repelling periodic point.

2. CURVE INTERSECTIONS REVEAL SYSTEM DYNAMICS

2.1 Roots and windows. In the bifurcation diagram, the gray haze of iterates contains fascinating gaps, or vertical strips of white space, usually called *windows*.

The windows are formed whenever the iteration abruptly changes from a wide spread of values to periodic oscillation. Each window can be associated with the lowest period of oscillation that occurs within that parameter range. A few windows are visible in Figure 1 but the Q-curves mark many windows that are too narrow to be visible.

A quick observation from Figure 2 is that the curve Q_3 crosses the axis in the region usually called the window of period 3. Likewise, Q_5 and Q_6 cross in windows of period 5 and 6, respectively. The natural conjecture is true: whenever a Q-curve meets the axis, it marks a periodic window in the diagram (with period of the lowest numbered Q-curve that crosses there) and all windows are marked in this way. Actually, each root is a parameter value with a superattracting periodic cycle. Recall that p is a *periodic point of period* n if $f_c^n(p) = p$ but $f_c^j(p) \neq p$ for $0 < j < n$. This p is *repelling* if $|(f_c^n)'(p)| > 1$, *attracting* if $|(f_c^n)'(p)| < 1$, and *superattracting* if $|(f_c^n)'(p)| = 0$.

Superattracting Root Theorem. *Let* $n \in \mathbb{N}$. *The parameter* r *satisfies* $Q_n(r) = 0$ *and* $Q_j(r) \neq 0$ *for* $0 < j < n$ *if and only if iteration of* $f_r(x)$ *has a superattracting periodic point of period* n.

Proof: Fix r. By definition, $Q_n(r) = f_r^n(0)$. Thus, $Q_n(r) = 0$ and $Q_j(r) \neq 0$ for $0 < j < n$ if and only if zero is a periodic point of period n. In general, a periodic point is superattracting if and only if a critical point belongs to the cycle. Indeed, by the chain rule, $(f^n)'(x_0) = \prod_{i=0}^{n-1} f'(x_i)$ and the product is 0 if and only if $f'(x_i) = 0$ for some i **[Br]**. Since zero is the only critical point of f_r, the theorem follows. ∎

Such parameter values do mark windows. Indeed, continuity implies that every parameter with superattracting period n is contained in an open interval of parameter values with attracting period n. Conversely, any maximal interval of attracting period n will contain a superattracting value. In fact, more is true: the derivative of the nth composition at a cycle point ranges from 1 to -1 across an interval of attracting period n. This statement follows from the more general complex result of Douady, Hubbard, and Sullivan (**[Br]**, p. 85) and can be observed in the diagrams of Devaney **[D2]**. A *window of period* n in the bifurcation diagram is an interval of attracting period n that is augmented on the left by a region of bounded iterates that we identify as an envelope in Section 3.2.

The roots of the Q-curves indicate many windows that are not visible in the bifurcation diagram (Figure 1) or even in zooms on the diagram **[D2]**. In Figure 3, the graph of Q_{11} alone marks a tremendous number of windows of period 11. Curve Q_{11} has no roots (except 0) to the right of -1.5, but the zoom shows how roots, and therefore windows, are concentrated near -2.

2.2 Roots and Mandelbrot buds. The Superattracting Root Theorem can be used in the complex c-plane to mark components of period n. This is analogous to marking windows on the c-axis. The polynomials $Q_n(c) = f_c^n(0)$ make sense for complex c-values, even though the term "curve" is not appropriate in this context. At complex roots where these polynomials are zero, the system has a superattracting periodic point of period n. Around each root, the region of attracting period n is a hyperbolic component of the Mandelbrot set **[Br]**, and appears as a bud sprouting off a larger component or as a cardioid of a (mini-)Mandelbrot set. We can color neighborhoods of the roots, where $|f_c^n(0)| < \varepsilon$, in order to show the

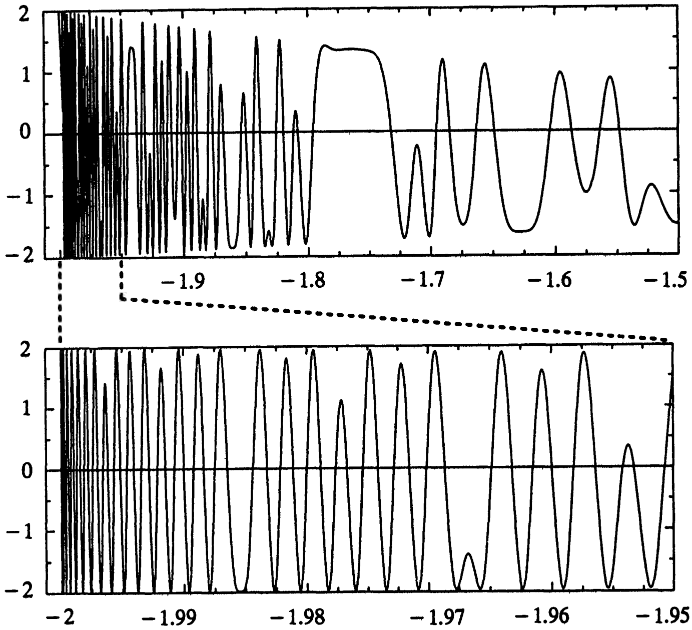


Figure 3. Every root of Q_{11} marks a period 11 window.

period of the corresponding components.

A program for plotting the Mandelbrot set is easily adapted to perform the color-coding of components. For each pixel c, the typical Mandelbrot set algorithm starts with $z_0 = 0$ and iterates f_c until $|z_n| > 2$, so that c is outside the Mandelbrot set, or n exceeds some maximum iteration, so that c is assumed to be in the Mandelbrot set. We simply add a third bail-out option: whenever $|z_n| < \varepsilon$, and $n \leq$ the number of colors to be used in period-labeling, stop and color the pixel with color n. Choose solid colors, for the original two stopping tests (as opposed to the usual practice of coloring according to n when $|z_n| > 2$). A fairly large ε, such as 0.2, is recommended to color "disks" that very roughly cover the buds that are being labelled. More importantly, it will label (the period of points in the main cardioid of) mini-Mandelbrot sets near the boundary that would not be visible and could even be smaller than a pixel.

This might be called a quick and dirty algorithm. It's quick because it's simple to program and actually provides an early exit for many pixels that would otherwise reach the maximum iteration. It's dirty because it's difficult to say anything rigorous about colored pixels for arbitrarily or aesthetically chosen ε's. It is usually easy to identify the component that goes with a colored disk, or vice versa, but colors can extend into adjacent components. Colored disks can be smaller or larger than the corresponding components, even for the same n and ε. Other algorithms, such as direct period detection from iteration or the spider algorithm, are more rigorous.

Two examples of the beautiful and intriguing color figures that are produced by this algorithm can be viewed through the World Wide Web at http://www.davidson.edu/academic/math/neidinger/road.html. Fascinating patterns appear in the sequence of periods found in the geometric patterns. Of course, zooms near the boundary are as entertaining as ever. A period-n bud, off

the main cardioid, is crowned by a decoration consisting of n spokes and each spoke contains a sequence of periods (covering cardioids of mini-Mandelbrot sets) from an equivalence class mod n.

2.3 The intersection dichotomy. We now leave the complex detour and return to the road map where every intersection is interesting. There are only two types of intersections between Q-curves and they imply different system dynamics for the corresponding parameter values. Figure 4 shows many intersections of the first eight Q-curves on $[-2, -1.8]$ and labels one of each type. A tangency occurs at a superattracting root r, and a non-tangent intersection occurs at a Misiurewicz point s, which has chaotic dynamics. While the implications of a root are relatively straight-forward, the implications of a Misiurewicz point rely on references in the literature.

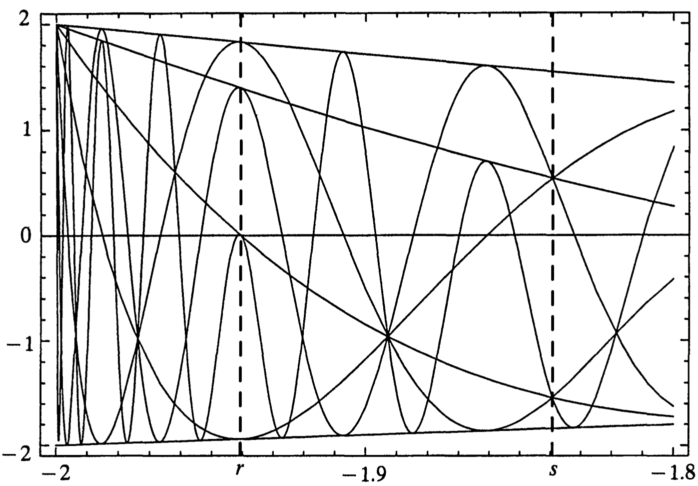


Figure 4. The two types of intersections: all curves tangent or none tangent, with periodic or chaotic dynamics, respectively.

A *Misiurewicz point* is a c-value where the orbit of the critical value zero is strictly pre-periodic, i.e., it eventually falls onto a periodic orbit not including zero. The proof that the corresponding dynamics is chaotic is very involved and uses the definition of chaos from ergodic theory (see references in the proof below). However, the following intuitive argument, from [CE], p. 34, yields some insight into the chaotic behavior: At a Misiurewicz point, the eventual periodic orbit of zero is repelling and, by Fatou's Theorem, no periodic attractor can exist. We now argue sensitive dependence on initial conditions, meaning that nearby points are eventually "pulled apart" through the iteration. This condition is hardest to fulfill for points near $x = 0$, since the derivative of $f_c(x)$ at zero is zero, and, hence, f_c contracts points near zero. But eventually zero itself is mapped onto a repelling periodic orbit. Thus, the orbits of points near zero may initially move closer (to the orbit of zero) but will eventually come near the repelling periodic orbit and be pushed away.

Intersection Dichotomy. *At any intersection of two Q-curves, the c-value is either*:

(1) *a root of some Q_n (let n be the smallest such positive number), where every intersection of Q-curves is a tangency at one of n distinct points (in equivalence*

classes mod n), and where the system f_c has a superattracting periodic point of period n, or

(2) *a Misiurewicz point, where no intersection of Q-curves is a tangency, and where the system f_c is chaotic.*

Proof: Let c be the value at the intersection of two Q-curves. By definition of Q_i, the orbit of zero is $Q_i(c)$. If c is not the root of any Q-curve, then the orbit never returns to zero but the intersection implies that the orbit is eventually periodic, so c is a Misiurewicz point. Thus, c is either a root or a Misiurewicz point.

The claims about system dynamics are found, in case (1), in the previous Superattracting Root Theorem and, in case (2), in results of M. Misiurewicz found in **[CE]** (p. 155 and supporting text). There, it is shown that if c is a Misiurewicz point, then f_c has an ergodic, invariant, probability measure that is absolutely continuous with respect to Lebesgue measure. This measure-theoretic condition is one way to define a *chaotic* dynamical system. The proof in **[CE]** applies to a class of functions, including $g = f_c/c$ for a Misiurewicz point $c \leq -0.5$. Ergodic theory techniques are used to construct an *invariant measure* ν (meaning $\nu(E) = \nu(g^{-1}(E))$ for every measurable set E) which is absolutely continuous. Such a measure is shown to be unique and, hence, ergodic. Readers unfamiliar with this characterization of chaos might want to consider how an attracting periodic orbit would prohibit the existence of an absolutely continuous invariant measure.

To prove the claims about tangency, observe that: an intersection and its tangency or non-tangency can be propagated to the next iterate (and hence all subsequent iterates) by the recurrence relations $Q_{i+1}(c) = (Q_i(c))^2 + c$ and $Q'_{i+1}(c) = 2Q_i(c)Q'_i(c) + 1$.

First, consider a root $c = r$ where $Q_n(r) = 0$ and no smaller numbered curve is zero at r. Then, $Q_{n+1}(r) = 0^2 + r = Q_1(r)$ and $Q'_{n+1}(r) = 2Q_n(r)Q'_n(r) + 1 = 1 = Q'_1(r)$. By the recurrence relations, $Q_{n+i}(r) = Q_i(r)$ and $Q'_{n+i}(r) = Q'_i(r)$ for all $i \geq 1$. So all these intersections are tangencies at one of the n points $Q_1(r), \dots, Q_n(r)$. There are no other intersections of Q-curves at r since these n points are distinct (otherwise, forward propagation would yield $Q_j(r) = Q_n(r) = 0$ for some $j < n$ which contradicts the assumption on n).

Now, assume that $c = s$ is a Misiurewicz point. Then, since zero is not itself periodic, s is not the root of any Q-curve. Let k be the smallest number such that Q_k intersects another Q-curve at s, and let n be the smallest number such that $Q_k(s) = Q_{k+n}(s)$. This is not a tangency, by the following argument. By the recurrence relation, $(Q_{k-1}(s))^2 = (Q_{k+n-1}(s))^2$ and, hence, $Q_{k-1}(s) = -Q_{k+n-1}(s)$. In particular, $k \geq 2$ (otherwise, the argument of the previous sentence would yield $0 = Q_{k+n-1}(s)$, which contradicts no root at s). Now, define $G(c) = Q_{k+n-1}(c) + Q_{k-1}(c)$. In the next paragraph, we will show that $G'(s) \neq 0$. Then

$$\begin{aligned} Q'_{k+n}(s) - Q'_k(s) &= 2Q_{k+n-1}(s)Q'_{k+n-1}(s) - 2Q_{k-1}(s)Q'_{k-1}(s) \\ &= -2Q_{k-1}(s)G'(s) \neq 0. \end{aligned}$$

We conclude that $Q'_{k+n}(s) \neq Q'_k(s)$, so the intersection is not a tangency. By the recurrence relations, $Q_{n+i}(s) = Q_i(s)$ and $Q'_{n+i}(s) \neq Q'_i(s)$ for all $i \geq k$. (We can propagate the non-tangency since s is not the root of any Q-curve.) Thus, all these intersections are non-tangencies at one of the n distinct points $Q_k(s), \dots, Q_{k+n-1}(s)$. There are no other intersections of Q-curves at s by the assumptions of "smallest" for k and n.

The following algebraic argument, from p. 333 of [DH], shows that $G'(s) \neq 0$: By the recurrence relation, $G'(s)/2 = Q_{k+n-2}(s)Q'_{k+n-2}(s) + Q_{k-2}(s)Q'_{k-2}(s) + 1$. To show that this expression cannot be zero, the rough idea is to show that the combination of Q and Q' values has a positive power of 2 as a type of factor and, thus, cannot be -1. Specifically, each rational number has a *2-adic valuation* given by the power of 2 in the natural factorization of the fraction. The field of rationals and the valuation can be extended to include the number s. Let A be the set of elements of the extension field with non-negative valuation and let m be the set of elements with positive valuation. Then, m is a maximal ideal in the ring A. Using obvious valuations, such as $1 \notin m$ and $2A \subset m$, we work with arithmetic mod m. Since s is the root of a monic polynomial with integer coefficients (i.e., an algebraic integer), s and all integer coefficient polynomials of s (including all Q and Q' values) are in A. In fact, every $Q'_i(s) \equiv 1 \mod m$, since $Q'_i(s) = 2Q_{i-1}(s)Q'_{i-1}(s) + 1$. Thus,

$$Q_{k+n-2}(s)Q'_{k+n-2}(s) + Q_{k-2}(s)Q'_{k-2}(s) \equiv Q_{k+n-2}(s) + Q_{k-2}(s) \mod m.$$

A property of s, from the previous paragraph, is that $Q_{k+n-1}(s) + Q_{k-1}(s) = 0$. By the recurrence, $Q^2_{k+n-2}(s) + Q^2_{k-2}(s) = -2s \in m$. It follows that $(Q_{k+n-2}(s) + Q_{k-2}(s))^2$ and, hence, $Q_{k+n-2}(s) + Q_{k-2}(s)$ are in m. By the preceding modular equivalence, $Q_{k+n-2}(s)Q'_{k+n-2}(s) + Q_{k-2}(s)Q'_{k-2}(s)$ is in m and cannot be -1 (which has valuation 0). Thus, $G'(s) \neq 0$. ∎

Now, one can literally see periodic and chaotic phenomena mixed throughout parameter ranges where the bifurcation diagram just shows a gray haze. Look at the curves to the left of the point marked r in Figure 4; there is superattracting periodic dynamics at every intersection with the axis, while there is chaotic dynamics at every non-tangent intersection of curves. Natural questions arise about the density and/or measure of the attracting periodic and chaotic parameters. The known answers are deep results in the field. Recently, Świątek showed that the union of the attracting periodic parameters is an open, dense set in the interval [S] and, yet, Jakobson's theorem shows that the set of chaotic parameters has positive Lebesgue measure [R]!

3. CURVE INEQUALITIES SHAPE THE DIAGRAM

3.1 Why curves appear in the bifurcation diagram. In the bifurcation diagram, Figure 1, the first few Q-curves appear as pathways that are well-traveled by later iterates. Once the iterates spread across a vertical range (for $c <$ the Feigenbaum point $-1.401155\ldots$, [C], p. 141), they seem to cluster around curves, particularly on one side of each curve. This density of iterates is especially clear in the histograms of [D4], p. 127, and [PJS], p. 632. Why do the Q-curves, formed by initial iterates of zero, emerge from these later iterates, even when $x_0 \neq 0$?

One way to explain this phenomenon is to show how iterates in a relatively wide band around the axis must map into narrow bands around the first few Q-curves. These bands are shown in Figure 5, between solid Q-curves and corresponding dashed curves, for c in $[-2, -1]$. The dashed curves are given by $P_n(c) = f_c^n(\varepsilon)$ (for $\varepsilon = 0.15$ in Figure 5), corresponding to $Q_n(c) = f_c^n(0)$. If any iterate x_i falls in the band of width 2ε around the c-axis, then the next iterate will be in the band of width ε^2 between Q_1 and P_1, which condenses iterates above Q_1. Specifically, if $0 \leq |x_i| \leq \varepsilon$, then, by squaring and adding c to each side, $Q_1(c) \leq x_{i+1} \leq P_1(c)$. Assuming $P_1(c) \leq 0$ and again applying f_c, yields $Q_2(c) \geq x_{i+2} \geq P_2(c)$. Since

$P_2(c) \geq 0$ for all $c < -1.4$, we can "propagate the bounds" to the next iterate, $Q_3(c) \geq x_{i+3} \geq P_3(c)$; so that the density of points will occur below Q_3. In the range of c-values where $Q_3(c) \leq 0$, the next iterate x_{i+4} will fall above Q_4 but, in the c-range where $P_3(c) \geq 0$, it falls below Q_4. The bands should appear condensed when the width $|Q_i(c) - P_i(c)|$ is smaller than the initial 2ε, which happens for low i and small ε, though the width can grow as c approaches -2. We can estimate $|Q_{i+1} - P_{i+1}| = |Q_i^2 - P_i^2| \leq (|Q_i| + |P_i|)|Q_i - P_i| \leq 4|Q_i - P_i|$. By induction, $|Q_{i+1} - P_{i+1}| \leq 4^i\varepsilon^2$. This phenomenon emphasizes one side of low-order Q-curves (lower should be darker); it can be observed in the bifurcation diagram in Figure 1.

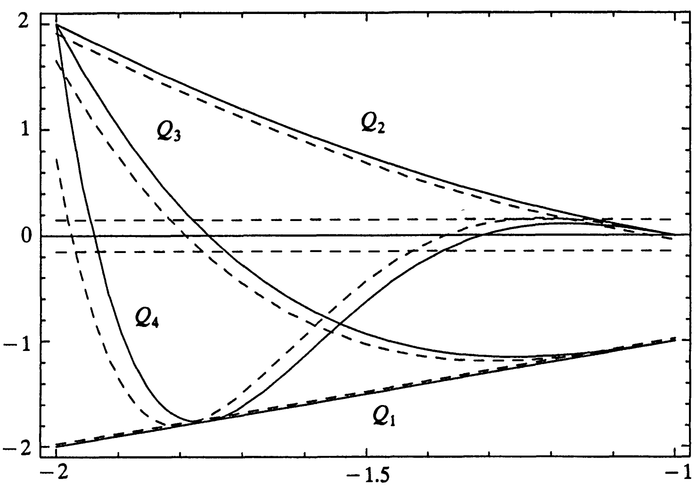


Figure 5. Iterates near the axis are mapped into concentrated bands around curves.

A geometrical view of the dynamical system can explain why the whole region below Q_3 is darker than the region above it, as c ranges between about -1.8 to -1.4. The bifurcation diagram on $[-2, -1]$ can be divided into two regions, in the (c, x)-plane, depending on whether or not $|x| \leq Q_2(c)$. If (c, x) satisfies $0 \leq |x| \leq Q_2(c)$, applying f_c yields $Q_1(c) \leq f_c(x) \leq Q_3(c)$; i.e., it maps the symmetric region by folding it over along the axis, stretching the axis down to Q_1, and stretching Q_2 down to Q_3. The remaining portion of the bifurcation diagram satisfies $Q_1(c) \leq x \leq -Q_2(c)$, so that applying f_c yields $Q_2(c) \geq f_c(x) \geq Q_3(c)$; flipping it over and filling in above Q_3. In the range between about -1.8 to -1.4, this motion condenses the first region and expands the second.

3.2 Envelopes bound copies of the diagram. On the road map given by curves on top of the bifurcation diagram (Figure 2), Q-curves form boundary lines separating regions that contain iterates from those that don't. In fact, these curves bound the shape and identify the location of all embedded, self-similar copies of the diagram. In general, whenever a curve Q_n crosses the axis, the curves Q_n and Q_{2n} bound a self-similar copy of the bifurcation diagram consisting of every nth iterate. The other iterates fall in similar envelopes, also bounded by Q-curves.

Figure 6 shows three such sets of envelopes. First, the entire diagram is bounded by Q_1 below and Q_2 above. Second, all the even iterates are bounded between Q_2 and Q_4, from the root of Q_2 to the point where Q_4 meets $-Q_2$. In

this parameter range, the odd iterates fall between Q_1 and Q_3. Finally, a quarter of the iterates are bounded between Q_4 and Q_8, from the root of Q_4 to the point where Q_8 meets $-Q_4$. Four envelopes can be seen in this parameter range.

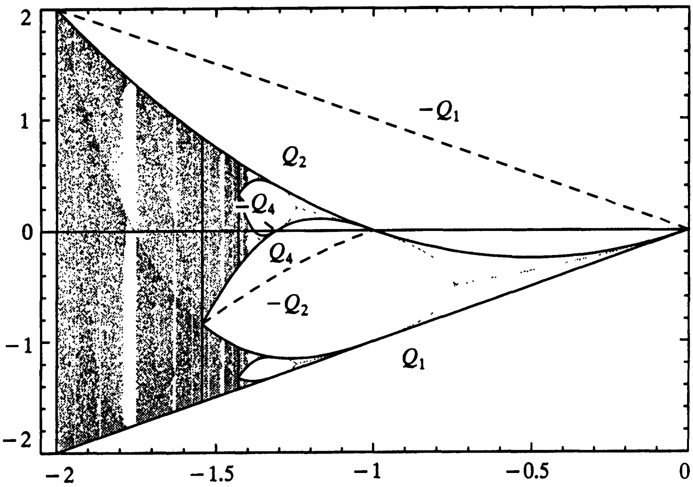


Figure 6. Q-curves bound nested copies of the diagram.

Figure 7 zooms in on the period three window, around the axis, to show an envelope formed when Q_3 crosses the axis. The envelope abruptly ends at the Misiurewicz point where $-Q_3$ meets Q_6. Within this range, every third iterate falls between Q_3 and Q_6, while the other iterates are constrained to the region between Q_1 and Q_4 or the region between Q_2 and Q_5. This containment is being observed when people speak of a "period 3 window." This is different from the interval of attracting period 3. Actually, if the envelope is on the interval (s, r) and the interval of attracting period 3 is (a, b), then the window is the union (s, b). (The right endpoint, b, of the attracting interval is a point not described by Q-curve intersections.) Each of the tiny, nested windows seen in Figure 7 contains even smaller envelopes of Q-curves.

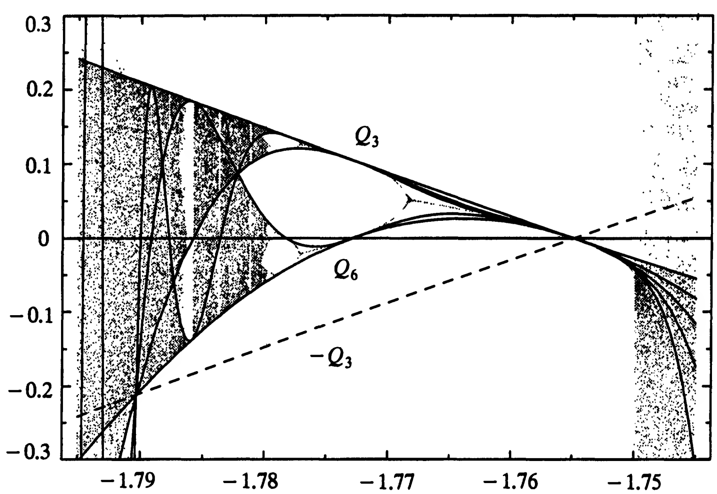


Figure 7. A zoom into the period 3 window shows an envelope of Q-curves.

The following theorem describes all such envelopes. The phrase x *lies between* y *and* z means $y \leq x \leq z$ or $z \leq x \leq y$.

Envelope Theorem. *If* $Q_n(r) = 0$ *and* $Q_j(r) \neq 0$ *for* $0 < j < n$, *let* $s = \max\{c < r\colon\ Q_{2n}(c) = -Q_n(c)\}$. *If* $c \in [s, r]$ *and* $|x_0| \leq |Q_n(c)|$, *then* x_{kn+m} *lies between* $\mathrm{Q_m}$ *and* $\mathrm{Q_{m+n}}$ *for all* $m \in \{1, 2, \dots, n\}$ *and* $k \in \{0, 1, 2, \dots\}$. *In particular,* Q_{kn+m} *lies between* Q_m *and* Q_{m+n} *on* $[s, r]$. *The regions between these pairs of curves,* Q_m *and* Q_{m+n} *for each* $m \in \{1, 2, \dots, n\}$, *are disjoint from each other on* $(s, r]$. *For later application, we also note that, if* $m \neq n$, *the region between* Q_m *and* Q_{m+n} *never meets the axis on* $[s, r]$.

To prove the Envelope Theorem, we "propagate the bounds" from one iterate to the next. However, this requires either that the iterate must fall between $-Q_n$ and $+Q_n$ or that both the upper and lower bounds must have the same sign. This will be reduced to the requirement that $|Q_{2n}| < |Q_n|$, a fact that is fairly obvious in the graphs. This and other essential properties of the "horn shape" formed by the graphs of Q_{2n} and Q_n (see Figures 6 and 7) are summarized in the following lemma. The claims following "moreover" are used in the next section.

Horn Lemma. *If* $Q_n(r) = 0$ *and* $Q_j(r) \neq 0$ *for* $0 < j < n$, *then there exists* $s < r$ *such that* $Q_{2n}(s) = -Q_n(s)$ *and* $|Q_{2n}| < |Q_n|$ *on* (s, r). *Moreover, there exists* $r' \in (s, r)$ *such that* $Q_{2n}(r') = 0$, $Q_{2n}Q_n < 0$ *on* $[s, r')$ *and* $Q_{2n}Q_n > 0$ *on* (r', r).

This lemma is deceptively innocent-looking—it embodies deep properties of the quadratic map that we will not prove. The "horn shape" is a real-parameter consequence of Douady and Hubbard's mini-Mandelbrot sets for the complex parameter **[DH]**; they prove that only complete (not partial) copies of the set occur. The horn shape is also equivalent to a pattern observed in graphical analysis (used to produce web-diagrams), as described in the following paragraph.

Consider the graph of $y = f_c^n(x)$ near the origin in the xy-plane. If $Q_n(r) = 0$, zero is a superattracting periodic point of period n for f_r, so that $y = f_r^n(x)$ has an extreme value at the origin. In fact, for a range of c-values near r, the graph resembles a parabola in a small region around the origin. We examine the concave-upward case shown in Figure 8. As c decreases from r to a value that we call s, the vertex moves monotonically downward until it becomes as large, in absolute value, as the coordinates of the intersection with the diagonal in the first quadrant. This diagonal intersection is a repelling point that forms the corner of a square of trapped iterates, as in **[D1]**, p. 132. The properties of the Horn Lemma can now be deduced from this family of graphs. On each graph of $y = f_c^n(x)$, the y-intercept is $Q_n(c) = f_c^n(0)$. Now, $Q_{2n}(c) = f_c^n(Q_n(c))$ is visualized by taking the y-intercept horizontally to the diagonal and then vertically to the curve. In Figure 8, the resulting values of $Q_{2n}(c)$ are connected, creating a curve parameterized by $x = Q_n(c)$ and $y = Q_{2n}(c)$ for c in $[s, r]$. This parametric curve displays all of the properties stated in the Horn Lemma. This is not a proof, but merely shows that the lemma is equivalent to the claim that the parabola-shape always moves monotonically until reaching the edge of the trapping box. Graphical analysis also shows the dynamical significance of the parameter s—beyond s, the parabola pokes out of the box and the containment (or envelope) of iterates ends. We now return to the c-axis for the proof of the containment between Q-curves.

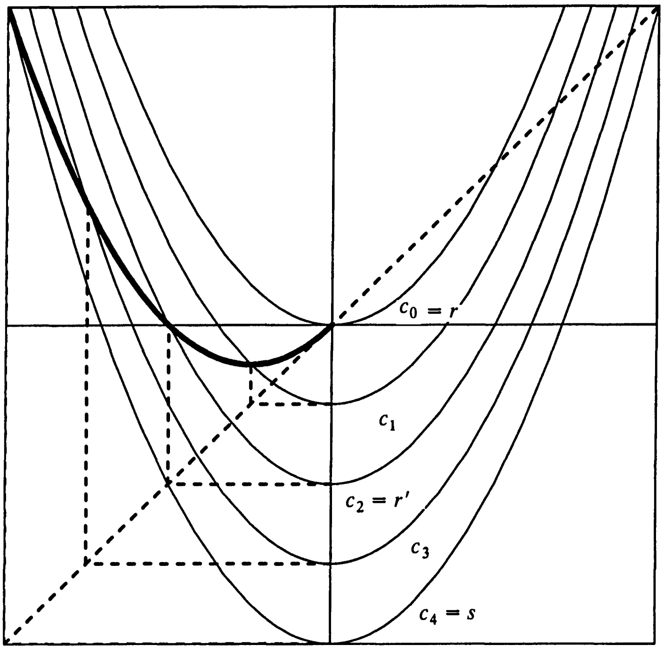


Figure 8. Five graphs of $y = f_c^n(x)$ lead to the Horn Lemma as shown in the thick curve parameterized by $x = Q_n(c)$ and $y = Q_{2n}(c)$.

Proof of the Envelope Theorem: (Assuming the Horn Lemma.) Fix r and n such that $Q_n(r) = 0$ and $Q_j(r) \neq 0$ for $0 < j < n$. By the Horn Lemma, let $s < r$ such that $Q_{2n}(s) = -Q_n(s)$ and $|Q_{2n}(c)| < |Q_n(c)|$ for $c \in (s, r)$. Clearly, $s = \max\{c < r: Q_{2n}(c) = -Q_n(c)\}$.

Fix any $c \in [s, r]$ and consider the orbit from any $|x_0| \leq |Q_n(c)|$. Apply f_c to $0 \leq |x_0| \leq |Q_n(c)|$ to get $Q_1(c) \leq x_1 \leq Q_{n+1}(c)$. The containment asserted by the theorem continues because of two propagation principles. (1) If x lies between Q_m and Q_{m+n} for some $m \in \{1, 2, \dots, n-1\}$, then $f_c(x)$ lies between Q_{m+1} and Q_{m+1+n}. (2) If x lies between Q_n and Q_{2n}, then $f_c(x)$ lies between Q_1 and Q_{1+n}. The second principle uses the fact that $|Q_{2n}(c)| \leq |Q_n(c)|$ for $c \in [s, r]$. If x lies between Q_n and Q_{2n}, then $0 \leq |x| \leq |Q_n(c)|$ and, again by applying f_c, $Q_1(c) \leq f_c(x) \leq Q_{n+1}(c)$. To prove the first principle, suppose x lies between Q_m and Q_{m+n}. One possibility (out of two) is that $Q_m(c) \leq x \leq Q_{m+n}(c)$. In the following paragraph, we show that $Q_m(c)$ and $Q_{m+n}(c)$ agree in sign. If they are both positive, then $Q_{m+1}(c) \leq f_c(x) \leq Q_{m+1+n}(c)$. If they are both negative, then $Q_{m+1}(c) \geq f_c(x) \geq Q_{m+1+n}(c)$. In either case, $f_c(x)$ lies between Q_{m+1} and Q_{m+1+n}. The other possibility is that the original containment was $Q_{m+n}(c) \leq x \leq Q_m(c)$, and the argument is identical. To complete the proof of containment, it remains to show that Q_m and Q_{m+n} are nonzero and agree in sign on $[s, r]$. We will show a bit more as a bonus.

We claim that the curves $Q_1, Q_2, \dots, Q_{2n-1}$ never meet the axis in (s, r) and that none of the curves $Q_1, Q_2, \dots, Q_{2n}$ intersect in (s, r). At the root r, the curves pair up, mod n, at the n distinct points (by the Intersection Dichotomy). Thus, the claim implies that the regions between these pairs of curves, Q_m and Q_{m+n} for each $m \in \{1, 2, \dots, n\}$, are disjoint from each other on $(s, r]$. The claim also implies that, for each $m \in \{1, 2, \dots, n-1\}$, the pair Q_m and Q_{m+n} are nonzero

and agree in sign throughout $(s, r]$. Since s is a Misiurewicz point, it can't be a root of Q_m or Q_{m+n}. So, in fact, Q_m and Q_{m+n} are nonzero and agree in sign throughout $[s, r]$. We prove the claim by contradiction. Let B be the set of counterexamples, i.e., $B = \{c \in (s, r)\colon Q_i(c) = Q_j(c)$ for some $i < j \leq 2n$, or $0 = Q_j(c)$ for some $j < 2n\}$, and suppose B is not empty. B is finite, since every Q_i is a distinct polynomial, so we can let $b = \max(B)$. Since $Q_i(b) = Q_j(b)$ or $0 = Q_j(b)$, propagate (apply f_b) until $Q_k(b) = Q_{2n}(b)$ for some $k < 2n$. Now, consider the continuous curves $|Q_k(c)|$ and $|Q_n(c)|$ on (b, r). At b, $|Q_k(b)| = |Q_{2n}(b)| < |Q_n(b)|$. By this inequality, $k \neq n$. At r, $|Q_k(r)| > 0 = |Q_n(r)|$. By the Intermediate Value Theorem, there exists some d in (b, r), such that $|Q_k(d)| = |Q_n(d)|$. Propagate this equality once to get $Q_{k+1}(d) = Q_{n+1}(d)$. Thus $d \in B$, which contradicts the assumed maximality of b. ■

The Envelope Theorem's hypothesis that $|x_0| \leq |Q_n(c)|$ is necessary to have containment for all c in $[s, r]$ but, in practice, is not needed for most c-values. The bifurcation diagram, in the background of Figure 7, was created with $x_0 = 0$ to satisfy this hypothesis, which essentially requires that the iteration start inside an envelope. For other x_0 values, the diagram looks almost the same but there can be isolated exceptions where iterates leak out of the envelopes. Outside the envelopes, an x_0 value can be, for instance, in a repelling periodic orbit for various values of c (with different periods) in the range of the envelope. At such c values, the iteration never enters the envelopes; nearby, it enters only after many iterations. Still, for most c-values, the iteration is rapidly attracted down into the envelopes. In fact, for a given c in the range of an envelope, the set of those values of x_0 with bounded orbits that are not attracted into the envelopes is a set of Lebesque measure zero.

3.3 Šarkovskiĭ's ordering. The Q_n curves also reveal the pattern known as Šarkovskiĭ's ordering. This ordering of $\mathbb{N}$ is as follows:

$$3 \rhd 5 \rhd 7 \rhd \cdots \rhd 2 \cdot 3 \rhd 2 \cdot 5 \rhd 2 \cdot 7 \rhd \cdots \rhd 2^2 \cdot 3 \rhd 2^2 \cdot 5 \rhd \cdots \rhd 2^3 \cdot 3 \rhd 2^3 \cdot 5 \rhd \cdots$$
$$\rhd \cdots \rhd 2^4 \rhd 2^3 \rhd 2^2 \rhd 2 \rhd 1.$$

Šarkovskiĭ's theorem **[D2]** applies to any continuous $F\colon \mathbb{R} \to \mathbb{R}$ and states that if F has a periodic point of period n and if $n \rhd m$, then F also has a periodic point of period m. Although this theorem applies to f_c for a fixed parameter c, it's not immediately clear how this relates to Q_n across a range of c-values. We will not use Šarkovskiĭ's theorem but will establish a relationship by directly studying the Q-curves.

Root Ordering Theorem. *Define $r_n = \max\{r\colon\ Q_n(r) = 0$ and $Q_j(r) \neq 0$ for $0 < j < n\}$. Then $r_n < r_m$ if and only if $n \rhd m$.*

Proof: Every r_n exists since, for $n > 2$, $Q_n(-2) = 2$ and $Q_n < Q_2 < 0$ on (r_2, r_1). By bifurcation theory (or the Horn Lemma and the Envelope Theorem), $\cdots < r_8 < r_4 < r_2 < r_1$ and if i is not a power of 2, then r_i is to the left of this list. For each power of two, $p = 2^j$, there is a corresponding $s_p < r_p$ described by the Horn Lemma. Also, $s_p < s_{2p} < r_{2p} < r_p$. These inequalities follow from the Horn Lemma; s_p can't be in (s_{2p}, r_{2p}) since on this interval $|Q_{4p}| < |Q_{2p}|$ and yet, at s_p, it can be argued that $Q_{4p}(s_p) = Q_{2p}(s_p)$. We conclude that $s_1 < s_2 < s_4 < s_8 < \cdots < r_8 < r_4 < r_2 < r_1$. To complete the proof, we now prove that, for each power of two $p = 2^j$, $s_p < r_{3p} < r_{5p} < r_{7p} < \cdots < s_{2p}$. The reader can use Figure 9 to follow along for $p = 1$.

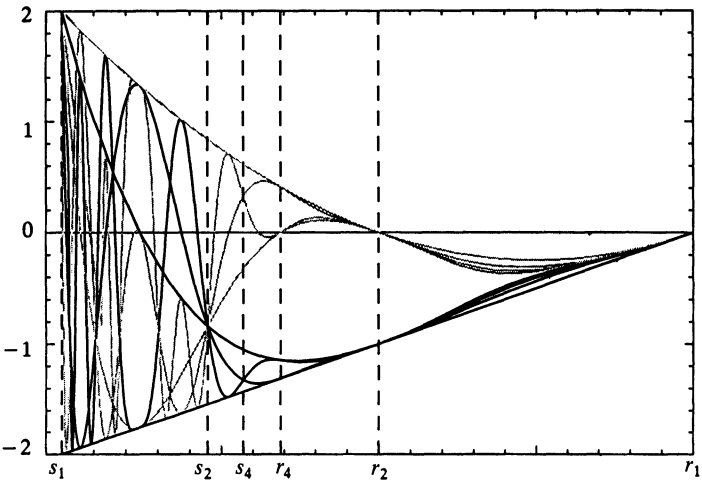


Figure 9. Šarkovskiĭ's ordering appears as odd curves (shown in black) cross the axis between s_1 and s_2.

Let $k \in \{3, 5, 7, \dots\}$. First, we have already argued that bifurcation theory (or the Horn Lemma and the Envelope Theorem) ensures that $r_{kp} < r_{2p}$. On $[s_{2p}, r_{2p}]$, the Envelope Theorem for $n = 2p$ says that Q_{kp} lies between Q_p and Q_{3p} and the region never meets the axis. Thus, $r_{kp} < s_{2p}$. We now establish that the roots do occur in the interval (s_p, s_{2p}). We will repeatedly use the fact that, at s_{2p}, the sign of each Q_{kp} agrees with the sign of Q_p, a fact implied by the observed containment of Q_{kp} curves.

By the Horn Lemma for $n = p$, Q_p and Q_{2p} have opposite signs throughout $[s_p, r_{2p})$, in particular throughout $[s_p, s_{2p}]$. At s_p, $-Q_p(s_p) = Q_{2p}(s_p)$ and iteration yields $Q_{2p}(s_p) = Q_{3p}(s_p)$. Thus, the sign of Q_{3p} changes from the sign of Q_{2p} at s_p to the sign of Q_p at s_{2p}. By the Intermediate Value Theorem, there exists a largest root r_{3p} of Q_{3p} in (s_p, s_{2p}). Now, at r_{3p}, $Q_{5p}(r_{3p}) = Q_{2p}(r_{3p})$. The sign of Q_{5p} changes from the sign of Q_{2p} at r_{3p} to the sign of Q_p at s_{2p}. By the Intermediate Value Theorem, there exists a largest root r_{5p} of Q_{5p} in (r_{3p}, s_{2p}). Again, at r_{5p}, $Q_{7p}(r_{5p}) = Q_{2p}(r_{5p})$, so that the sign of Q_{7p} changes on (r_{5p}, s_{2p}). By induction, $s_p < r_{3p} < r_{5p} < r_{7p} < \cdots < s_{2p}$. ■

4. CONCLUSION. Many of the most intriguing features along the road to chaos are clarified by looking at the road map showing Q-curves. For years, people have studied the wonders of the bifurcation diagram: there appears to be a region of chaos that is interrupted by windows of periodic attraction and that has shadowy curves from varying density of iterates; the entire diagram appears in smaller, similar copies nested within the diagram. Each of these phenomena is better understood if one also studies the Q-curves. The roots of Q-curves show that periodic windows are much more prevalent than the diagram suggests. On the other hand, the non-tangent intersections of Q-curves point out at least some points where one can confidently say that the system is chaotic. The Q-curves are the curves that appear in the diagram, both as the shadowy curves of higher density and as the boundaries of the diagram and the nested copies. Indeed, any root of a Q-curve begins, and a corresponding intersection ends, a nested copy of the diagram. Even Šarkovskiĭ's ordering and a color-coded (according to period) Mandelbrot set arise from this study of Q-curves.

Most of our proofs use only elementary mathematics and most of our arguments remain on the parameter axis, rather than switch to web diagrams on the graph of the quadratic. The key ideas are very accessible—periodic attraction at roots uses the chain rule, tangencies and envelopes use elementary algebra on equalities and inequalities, and Šarkovskiĭ's ordering uses the Intermediate Value Theorem. However, the properties of Misiurewicz points and the Horn Lemma do rely on more advanced work in ergodic theory, algebra, and complex analysis. There is ample graphical evidence for the Horn Lemma, which plays a crucial role in the Envelope Theorem and Šarkovskiĭ's ordering.

ACKNOWLEDGMENTS. We thank R. L. Devaney and many other colleagues and students for helpful conversations while developing this paper.

REFERENCES

[Br] B. Branner, The Mandelbrot Set, in *Chaos and Fractals*, Devaney and Keen, eds. Amer. Math. Soc., Providence, RI, 1989, pp. 75–105.

[C] R. M. Crownover, *Introduction to Fractals and Chaos*, Jones & Bartlett, Boston, 1995.

[CE] P. Collet and J-P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems*, Birkhäuser, Boston, 1980.

[D1] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Second Ed., Addison-Wesley, Redwood City, CA, 1989.

[D2] R. L. Devaney, Dynamics of Simple Maps, in *Chaos and Fractals*, Devaney and Keen, eds. Amer. Math. Soc., Providence, RI, 1989, pp. 1–24.

[D3] R. L. Devaney, The Orbit Diagram and the Mandelbrot Set, *College Math. J.*, 22 (1991), 23–38.

[D4] R. L. Devaney, *A First Course in Chaotic Dynamical Systems*, Addison-Wesley, Reading, MA, 1992.

[DH] A. Douady and J. H. Hubbard, On the Dynamics of Polynomial-like Mappings, *Ann. Scient. Ec. Norm. Sup.*, 4 serie, 18 (1985) 287–343.

[PJS] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals*, Springer-Verlag, New York, 1992.

[R] M. R. Rychlik, Another Proof of Jakobson's Theorem and Related Results, *Ergodic Theory and Dynamical Systems*, 8 (1988) 93–109.

[S] G. Świątek, Hyperbolicity is Dense in the Real Quadratic Family, preprint.

Department of Mathematics
Davidson College
P.O. Box 1719
Davidson, NC 28036
rineidinger@davidson.edu

Forsyth County Management Info. Services
Room 604—Hall of Justice
200 N. Main Street
Winston-Salem, NC 27101
annen@zeus.co.forsyth.nc.us

Iteration of Quaternion Functions

Stephen Bedding and Keith Briggs

1. QUATERNIONS. Much of mathematics is carried out using the real ($\mathbb{R}$) or complex ($\mathbb{C}$) number systems. These two systems are the simplest examples of a larger category known as *division algebras*. A division algebra is a system of elements along with rules for addition and multiplication, in which every non-zero element has a multiplicative inverse. The concept of a quaternion arises when we attempt to extend the idea of a complex number by introducing an extra imaginary unit. Starting with the elements $\{1, \mathbf{i}\}$, such that $\mathbf{i}^2 = -1$, we append a new element $\mathbf{j}$ satisfying $\mathbf{j}^2 = -1$, $\mathbf{j} \neq \mathbf{i}$. The question immediately arises as to what results from forming the products $\mathbf{ij}$ or $\mathbf{ji}$. If we insist that the new system should be a division algebra, then it is straightforward to show (using consistency arguments) that these products cannot be proportional to 1, $\mathbf{i}$, or $\mathbf{j}$, and so must be something new, $\mathbf{k}$ say. Similar consistency arguments lead to the following multiplication rules

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1 \tag{1}$$

$$\mathbf{ij} = -\mathbf{ji} = \mathbf{k} \tag{2}$$

$$\mathbf{ik} = -\mathbf{ki} = -\mathbf{j} \tag{3}$$

$$\mathbf{jk} = -\mathbf{kj} = \mathbf{i}. \tag{4}$$

These equations define the unique 4-dimensional division algebra. This set with the rules (1–4) generates one of the two 4-dimensional *Clifford algebras*; the other set is not a division algebra.

A member of the division algebra defined by these equations is known as a quaternion, and has the form

$$q = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k} \tag{5}$$

with $q_0, q_1, q_2, q_3 \in \mathbb{R}$. The space of quaternions will be designated $\mathbb{H}$, after its discoverer Sir William Rowan Hamilton. It follows from the multiplication rules that quaternion multiplication is associative:

$$a(bc) = (ab)c \tag{6}$$

for any three quaternions a, b, c. However, $\mathbb{H}$ is not commutative; in general, $ab \neq ba$.

Many of the elementary properties associated with complex numbers have a natural generalization to the quaternions. Thus the Argand plane is replaced with a set of four mutually orthogonal directions representing 1, $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, which is of course impossible to visualize or draw in its entirety. Further, the quaternion q defined in equation (5) has a real and a purely imaginary part (the latter will be

referred to as the *pure* part)

$$\mathrm{Re}(q) = q_0 \tag{7}$$

$$\mathrm{Im}(q) \equiv \mathsf{Pu}(q) = \mathbf{i}q_1 + \mathbf{j}q_2 + \mathbf{k}q_3. \tag{8}$$

The *quaternionic conjugate* $\bar{q}$ is now obtained by changing the sign of the pure part:

$$\bar{q} = q_0 - \mathbf{i}q_1 - \mathbf{j}q_2 - \mathbf{k}q_3. \tag{9}$$

The *modulus* (or length) of q is a natural generalization of Pythagoras' theorem to 4 dimensions:

$$\|q\| = \left(q_0^2 + q_1^2 + q_2^2 + q_3^2\right)^{\frac{1}{2}} = (q\bar{q})^{\frac{1}{2}} = (\bar{q}q)^{\frac{1}{2}}. \tag{10}$$

Thus for $q \neq 0$,

$$q^{-1} = \frac{\bar{q}}{\|q\|^2}. \tag{11}$$

A *unit quaternion* is one for which $\|q\|^2 = q\bar{q} = \bar{q}q = 1$.

The space $\mathbb{H}$ contains an infinite number of two-dimensional subspaces isomorphic to $\mathbb{C}$. Each such subspace is spanned by the real unit of $\mathbb{H}$ along with any nonzero combination of the three imaginary units **i**, **j**, and **k**. The simplest examples are the orthogonal planes spanned by $\{1, \mathbf{i}\}$, $\{1, \mathbf{j}\}$, and $\{1, \mathbf{k}\}$. From equation (10) we see that the components of a unit quaternion satisfy

$$q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1. \tag{12}$$

This equation defines the surface of a 3-dimensional hypersphere S^3 of unit radius. Thus, the unit quaternions lie on the surface of such a 3-sphere. In a similar fashion, the unit pure quaternions (for which $q_0 = 0$ and $q_1^2 + q_2^2 + q_3^2 = 1$), define a 2-sphere of radius 1.

We conclude this section by examining one rather important difference between quaternions and complex numbers. If we consider sums of the form $\Sigma_{i=0}^{n} a_i z^i$ (with $z = x + \mathbf{i}y$), we find that the polynomial functions so constructed form a *proper subset* of all polynomial functions of two real variables x and y. That is, there exist complex polynomial functions of two real variables x and y that cannot be obtained from polynomial functions of $z = x + \mathbf{i}y$. The case of quaternions and four real variables is different. A quaternionic-valued polynomial in four real variables q_0, q_1, q_2, q_3 can always be written as a sum of terms of the form $a_0 q a_1 q \dots a_{n-1} q a_n$ for various choices of n and a_i. The reason for this is evident when one notices that each individual component of q can be written in such a way. For example,

$$q_0 = \tfrac{1}{4}(q - \mathbf{i}q\mathbf{i} - \mathbf{j}q\mathbf{j} - \mathbf{k}q\mathbf{k}). \tag{13}$$

The reader might like to derive the equivalent formulae for q_1, q_2, q_3, which may be found in **[1]**. The point is that in the complex case, there are no similar formulae that express x and y as linear functions of z alone; $\bar{z}$ is required also in this case. Both z and $\bar{z}$ are required because they are independent functions of x and y. On the other hand, using the formulae of **[1]**, $\bar{q}$ can be written explicitly as a function of q and is thus not independent of q.

For an excellent review of quaternionic algebra in general the reader may refer to **[2]**.

2. REGULARITY. For the complex numbers, the idea of a regular (or analytic) function (that is, one for which the rules of differentiation can be used unambigu-

ously everywhere) is closely related to the results described at the end of the last section concerning the independence of $z, \bar{z}$. Since the quaternion $\bar{q}$ is not independent of q, we can expect to see differences from the complex case in the treatment of differentiation of quaternions. The property of regularity is of crucial importance for the structure and study of complex Mandelbrot sets, so we must devote some time to it here.

Naïvely, one might hope to be able to extend the notion of differentiability directly to $\mathbb{H}$ by defining a function f to be *quaternion differentiable* (on the left, i.e., the factor h^{-1} appears on the left) at q, provided

$$\lim_{h \to 0} (h^{-1}(f(q+h) - f(q)))$$

exists. Unfortunately, the non-commutativity of $\mathbb{H}$ causes this definition to be far too restrictive, and the only functions that satisfy it have the form $f(q) = aq + b$, $(a, b \in \mathbb{H})$. The problem is easy to see in concrete form if the function $f(q) = q^2$ is examined. Since $(q+h)^2 = q^2 + qh + hq + h^2$, the limit reduces to $q + \lim_{h \to 0}(h^{-1}qh)$. A close look at the details of the limit (in components) reveals that it is ambiguous!

In view of this problem, it is necessary to seek alternative candidate definitions of regularity for quaternion functions. The problem was originally studied by Fueter [**3, 4, 5**], who generalized the so-called Cauchy-Riemann operator $\partial/\partial x \pm \mathbf{i}(\partial/\partial y)$ of the complex case to $\partial/\partial q_0 \pm (\mathbf{i}(\partial/\partial q_1) + \mathbf{j}(\partial/\partial q_2) + \mathbf{k}(\partial/\partial q_3))$. Such operators are now known as Cauchy-Riemann-Fueter (CRF) operators. One of these operators can be chosen to provide a useful definition of a regular function. Following Sudbery [**1**], we define

$$\bar{\partial}_l \equiv \frac{\partial}{\partial q_0} + \mathbf{i}\frac{\partial}{\partial q_1} + \mathbf{j}\frac{\partial}{\partial q_2} + \mathbf{k}\frac{\partial}{\partial q_3} \tag{14}$$

(regarded as acting from the left). Then a quaternion function f is said to be *left-regular at* q if

$$\bar{\partial}_l f(q) = 0. \tag{15}$$

A similar definition may be made for right-regularity. The term *monogenic* is also used for a regular quaternion function. This definition proves very useful in the study of quaternionic integral theorems and generalization of Cauchy's theorem, but suffers from the possible drawback that common functions of interest, such as polynomials and even the identity function itself, are not left-regular functions. For this reason, we include an alternative definition due to Rinehart [**6**] and Cullen [**7**]. We define the *Cullen differential operator* by

$$\partial_C \equiv \frac{1}{2}\left(\frac{\partial}{\partial q_0} + \frac{\mathsf{Pu}(q)}{r}\frac{\partial}{\partial r}\right), \tag{16}$$

where $r^2 \equiv q_1^2 + q_2^2 + q_3^2$. A function $f\colon q \mapsto f(q)$ is said to be *Cullen-regular* if

$$\partial_C f(q) = 0. \tag{17}$$

The variable r is a radial distance on a 2-sphere in the pure part of $\mathbb{H}$ and $\mathsf{Pu}(q)/r$ is then a Cullen-regular unit radial vector. ∂_C is a generalization of the usual complex differential operator $\partial/\partial z$. If we restrict q to any of the planes in which some pair of q_1, q_2, q_3 vanishes, then ∂_C and $\bar{\partial}_C$ are precisely the complex differential operators in that plane; thus it should not be surprising that, as in the

complex case, the Cullen definition of regularity includes all (real coefficient) polynomial functions of q as regular functions. This result follows easily from the observation that such polynomials can be expressed in the form

$$p(q) = f(q_0, r) + g(q_0, r)\frac{\mathsf{Pu}(q)}{r}. \tag{18}$$

Since $\mathsf{Pu}(q)/r$ commutes with ∂_C, the regularity follows by recognizing that the situation is identical to the case of Cauchy-Riemann equations for a complex polynomial.

As well as polynomial functions, Cullen-regular functions include all of the familiar functions (defined by their usual Taylor expansions) such as exponential, trigonometric, logarithmic, and rational functions. Real coefficients are assumed both for the overall function and for the argument of the function. The usual procedures for differentiation carry through, and we have

$$\begin{aligned} \partial_C q &= 0 \\ \bar{\partial}_C q &= 1 \\ \partial_C(q^n) &= 0 \\ \bar{\partial}_C(q^n) &= nq^{n-1} \\ \partial_C(f(q)) &= 0. \end{aligned} \tag{19}$$

We will make little use of the concept of Cullen-regularity; further uses of the unqualified term "regular" will refer to the definition in (15).

3. ITERATION THEORY. Iteration theory is the study of sequences obtained by repeatedly applying a given function (or *map*) to an initial point (or *seed*). We first make several definitions that we will use throughout this article. Here, the map f may be a function $\mathbb{X} \to \mathbb{X}$, where $\mathbb{X}$ is one of the spaces $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$.

Let $x_n = f \circ \cdots \circ f(x_0) \equiv f^{[n]}(x_0)$, $n = 0, 1, 2, 3, \ldots$, with x_0 a given seed. Then the sequence $\{x_n\}$, $n = 0, 1, 2, \ldots$ is called the *orbit* of x_0. A *fixed point* is a point x^* such that $x^* = f(x^*)$. An *n-cycle* is a fixed-point of the map $f^{[n]}$: $x_0 = f^{[n]}(x_0)$. Note that a 1-cycle is just a fixed point.

In applications, the most important n-cycles are those that are *stable*. This concept is easy to define in the cases of real and complex maps. Every cycle has a *stability*, defined by $f^{[n]\prime}(x^*)$, where x^* is any one of the points of the n-cycle. We say that an n-cycle is *stable* if its stability satisfies $|f^{[n]\prime}(x^*)| < 1$. Also important is the idea of a *superstable* n-cycle, which satisfies $f^{[n]\prime}(x^*) = 0$, and so a superstable n-cycle is of course stable. In the real and complex cases, we have the important theorem that a stable n-cycle has an attracting neighborhood.

These concepts need a little elaboration in the case of maps on the quaternions. Stability is no longer determined by the derivative of the map, but rather, we must compute the eigenvalues of the 4×4 Jacobian matrix, obtained by considering the map as a function $\mathbb{R}^4 \to \mathbb{R}^4$. A fixed point is then said to be stable if the absolute value of *all* the eigenvalues is less than one. To decide the stability of a quaternionic n-cycle, we must first compute the iterated map $f^{[n]}$, then its Jacobian matrix and eigenvalues.

4. THE MANDELBROT SET. Having established our terminology, we now consider a central idea of this article, the Mandelbrot set, or *Mandelset*. This set arises when we consider the particular family of complex quadratic maps $f_c(z) = c + z^2$.

Here z is a complex dynamical variable and c is a complex parameter. For a given value of the parameter c, we ask whether the map $c + z^2$ has a stable n-cycle, for any n. It is a deep theorem that if this occurs, there is no other stable m-cycle, for any $m \not\equiv 0 \pmod{n}$ **[8, 9]**. c is then an element of the Mandelset. However, such values of c do not constitute all of the Mandelset, because it happens that for certain c, f_c has a bounded orbit that does not approach an n-cycle for any n. Thus the Mandelset M is defined as $M = \{c \text{ such that } z_n \text{ is bounded}\}$, where $z_0 = 0$ and $z_{n+1} = f_c(z_n)$, $n = 0, 1, 2, \ldots$.

Before continuing, we will clarify two obvious questions: first, why do we use the particular form $f_c(z) = z^2 + c$, and not an arbitrary quadratic? The answer is that an arbitrary quadratic is *affine conjugate* to f_c. (An analytic, invertible change of coordinates ϕ is called a *conjugacy*, and if $f = \phi \circ g \circ \phi^{-1}$, f and g are said to be *conjugate* maps. An affine conjugacy has the form $z \mapsto az + b$, for constant a and b. Since it follows that $f^{[n]} = \phi \circ g^{[n]} \circ \phi^{-1}$, we see that conjugate maps are essentially equivalent with respect to iteration. They are just the 'same' map, viewed from different coordinate systems.) Second, why do we consider only the seed $z_0 = 0$? The following theorem justifies this:

Theorem. *If the map f_c has a stable cycle, the orbit with seed* 0 *is attracted to it.*

In summary, this means that if we want to know if a stable cycle exists, we can just start iterating with the seed 0. One of the following three things must happen, depending only on the value of c:

1. The orbit approaches a stable cycle, so that $c \in M$.
2. The orbit is bounded, but does not approach a stable cycle, and again $c \in M$ (very rare!).
3. The orbit diverges to infinity, so that $c \notin M$.

Thus the Mandelset classifies the behavior of the family of maps f_c. The most basic feature of the Mandelset is that it consists of components (typically approximately cardioids or disks) that meet at single points. These points, called $n$furcation values of c, are points at which a k-cycle loses stability, and a nk-cycle gains stability (see Figure 1). We will denote the stability of an orbit by ρ.

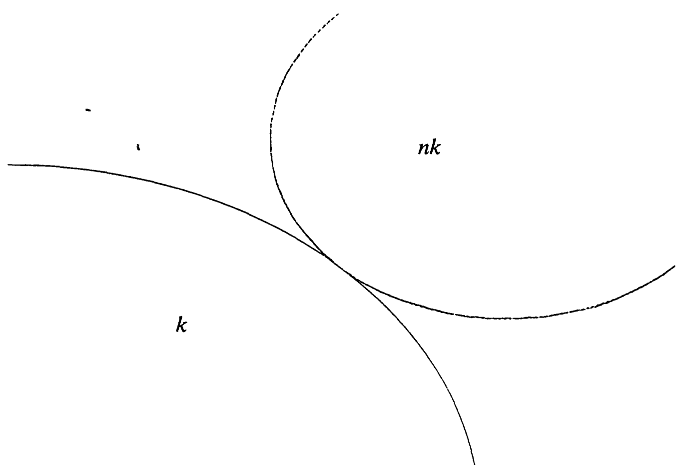


Figure 1. The structure of the Mandelset near an $n$furcation point.

At an $n$furcation value the stability of the common k and nk cycle must have modulus one. In fact, it takes the value $\exp(2\pi i m/n)$ for some integer $m < n$. Thus, the stability of the n-cycle is a nth root of unity. A detailed discussion of this $n$furcation behavior would make use of the fact that at an $n$furcation value of c, the map $z \mapsto c + z^2$ is locally analytically conjugate to a so-called *normal form* $z \mapsto \exp(2\pi i m/n)z + z^2$ **[8, 9]**. This fact plays an important role in our discussion.

We can now understand the local structure of the Mandelset. (We emphasize the word *local*. To understand the global structure is a far more difficult problem.) Thus, the picture in the parameter plane is locally two tangential curves, approximately arcs of circles. Note in particular, that the regions corresponding to stable k- and nk-cycles do not overlap. This behavior depends on the fact that we have an analytic complex map. We now proceed to study regular quaternion maps to see whether any of this picture is preserved.

5. REGULARLY ITERABLE LINEAR MAPS. We have explained how analyticity plays a crucial part in determining the highly intricate structure of the complex Mandelsets and gives us a powerful tool with which to delve into that structure. For this reason, it is important to investigate the iterative behavior of quaternion maps that possess the analogous property. In the complex case, once a regular map has been chosen, the property of regularity is preserved under iteration. This is not the case for quaternion maps, so if regularity is to play an equally important part, (and we see little hope of obtaining much insight without some such restriction), then we would expect that the interesting maps preserve regularity under iteration. We should therefore seek such maps. This turns out to be a highly non-trivial exercise that results in some very complicated calculations. Consequently, it is prudent to investigate the implications of the demand of regularity preservation for linear maps, before studying nonlinear maps.

By a regular linear (that is, real homogeneous of degree one) quaternion map we mean a map of the form

$$f(q) = (\mathbf{i}q + q\mathbf{i})a + (\mathbf{j}q + q\mathbf{j})b + (\mathbf{k}q + q\mathbf{k})c, \tag{20}$$

where a, b, c are constant quaternions. We define a *regularly iterable* map to be a regular map, all of whose iterates are regular. A simple application of the definition of regularity to a linear quaternion map shows that it is regular if and only if when considered as a map on $\mathbb{R}^4$ it has the following structure:

$$f(q) = -2Aq = -2\begin{bmatrix} a_1 + b_2 + c_3 & a_0 & b_0 & c_0 \\ -b_3 + c_2 - a_0 & a_1 & b_1 & c_1 \\ -c_1 + a_3 - b_0 & a_2 & b_2 & c_2 \\ -a_2 + b_1 - c_0 & a_3 & b_3 & c_3 \end{bmatrix}\begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix}, \tag{21}$$

where the 4×4 matrix A is defined by this equation. Higher iterates of the linear map are obtained by applying higher powers of A to q. To ensure that the second and third iterates are regular, we have to impose eight equations on the components of A, in order that A^2 and A^3 have the same structure as A itself. The Cayley-Hamilton theorem tells us that all powers of A above the third power can be written in terms of A, A^2, A^3 and the 4×4 identity matrix, with coefficients involving the trace invariants and determinant of A. Examination of the details of this relation for A^4 quickly reveals that we now need to impose only one further condition, $\det(A) = 0$, in order to have a regularly iterable map. More detailed explanation of this procedure is in **[10]**. The resulting equations are extremely complicated. In **[10]** we solved all of them either directly or by invoking symmetry

considerations, and we found all solutions. For all but one of the solution sets, the form of the matrix A is such that its eigenvalues are a complex conjugate pair and two zeroes. This implies that the linear map projects all initial quaternions into a 2-dimensional subspace of $\mathbb{H}$ and iterates within that subspace, (e.g., see **[11]**). Such maps are equivalent to linear maps in the complex plane and as such are uninteresting in a quaternionic context. The only exceptional case is represented by a matrix of the form

$$A = -c_3 \begin{pmatrix} 3 & 0 & \sqrt{3} & 0 \\ 0 & 2 & 0 & 0 \\ -\sqrt{3} & 0 & -1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \tag{22}$$

and the most general solution is obtained from this special case by a similarity transformation with a matrix M to MAM^{-1}, which preserves the eigenvalues of A. The 4×4 matrix M contains four parameters and is a general member of the group SO(1, 3). Further details and the general expression for M appear in **[10, 11]**.

The map has eigenvalues $(c_3, c_3, c_3, 0)$ and projects first into a three-dimensional subspace of $\mathbb{H}$. Further iterates then either fix a point in the subspace (for $c_3 = 1$) or shift the point along a ray to 0 or ∞, according to the modulus of c_3. In view of this behavior we have reluctantly concluded in **[10]** that there is nothing very interesting about regularly iterable linear maps, and that this conclusion probably extends also to nonlinear maps.

6. QUADRATIC QUATERNION MAPS. When pondering the subject of Mandelbrot sets in a quaternionic context, one of the first questions to pose is whether the complex Mandelset can be extended in a natural way into four dimensions, simply by iterating the function $f(q) = q^2 + c$, with quaternions q, c and seed 0. It is straightforward to write a short computer program to iterate this function and plot planar (two-dimensional) cross-sections for the parameter c showing the values of c for which the orbit is bounded. If such cross-sections be chosen to include the

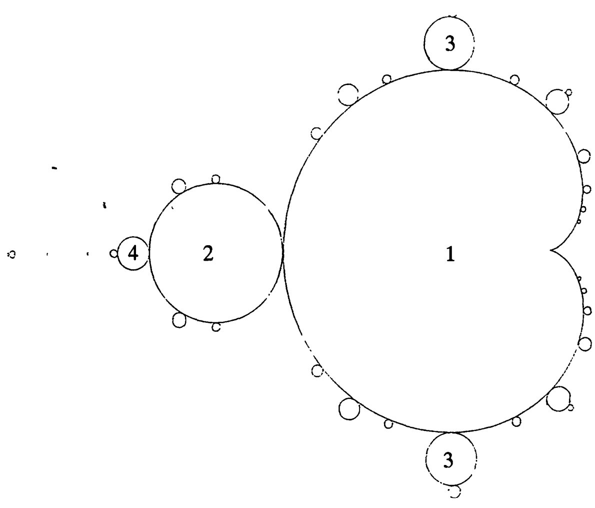


Figure 2. Quaternion Mandelbrot set with $c_2 = c_3 = 0$. The choice of c_2, c_3 means that this is just the usual complex Mandelset. Here, and in later figures, only the boundary of the Mandelset is shown. The integers are periods of the stable cycles existing in the respective components.

real c axis, then we see the familiar Mandelset (Figure 2) [12, 13]. Indeed, it is not hard to convince oneself that the regions of the set that symmetrically straddle the real axis are topological 3-spheres, while the regions that are offset from the real axis are slices through tubes circling the axis in the imaginary directions, with the same period throughout the tube. The results of some computer experiments supporting such a conclusion are shown in Figures 3–6.

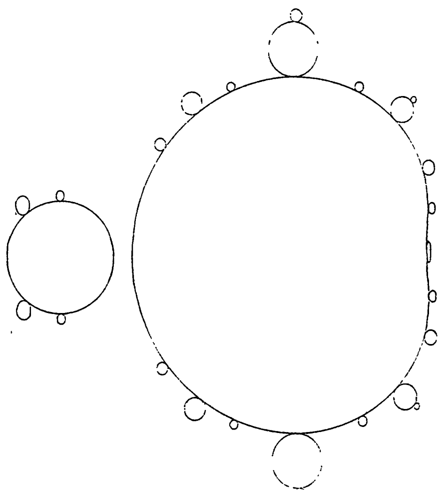


Figure 3. Quaternion Mandelbrot set with $c_2 = 0.15$, $c_3 = 0$, showing how the period two region on the left loses contact with the period one region on the right as we transport the cross sectional plane parallel to itself in the c_2 direction. c_2 and c_3 are equivalent by symmetry. This shows that the regions touch at only one point, like a pair of tangential billiard balls. Notice also that the cusp on the right of the period one region has become smooth. The small region that appears to have grown over the cusp position is a region of higher period located above and below the cusp on the edge of the standard Mandelset.

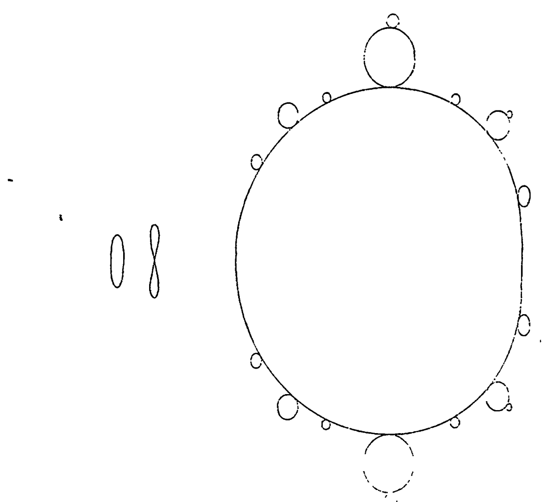


Figure 4. Quaternion Mandelbrot set with $c_2 = 0.25$, $c_3 = 0$. The regions are gradually shrinking as c_2 increases; in fact, the period two region has disappeared. The plane for this parameter choice also happens to cut two tubes of higher period, which now appear unobscured by the period two region.

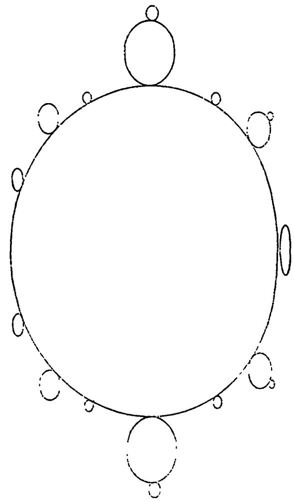


Figure 5. Quaternion Mandelbrot set with $c_2 = 0.35$, $c_3 = 0$. The region of period two and its higher period attached regions have disappeared altogether as our planar cross-section moved past its edge.

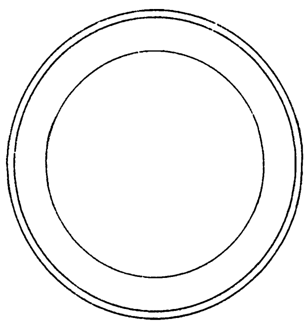


Figure 6. Here we see a cross-section of the quaternion Mandelbrot set in the c_2, c_4 plane with $c_1 = -0.125$, $c_3 = 0$. The value of c_1 was chosen so that the cross-section passes through two of the period three regions. In the *complex* Mandelset these two regions appear as circles on the top and bottom of the period one cardioid. This figure shows how these two originally-separate period three regions have become linked via the third and fourth dimensions available in the quaternion Mandelset, forming circular tubes in those extra dimensions. As we move out from the center of the concentric circles shown here, we pass from the period one region (which has spherical topology), through period three to period six, both of which are tubular regions.

Generally speaking, it has proved impossible to apply the ideas of Section 4 to the quaternionic case, because the detailed calculations there rely heavily on commutativity of the complex numbers. However, the particular case of $f(q) = q^2 + c$ is special and we may use the arguments from that section to gain some understanding of the bifurcation process. In earlier work **[11]** we have pointed out that the iterates of $q^2 + c$ starting from 0 remain in a plane defined by $(0, c, c^2)$, which is a complex subspace of $\mathbb{H}$. In this case, the detailed algebra mentioned in Section 4 takes place within the same complex subspace and there is no problem with commutativity. New developments occur at the point where one demands that the stability of the orbit (as defined in Section 3) should be an nth root of unity. Let us investigate this briefly.

Let the stability $\rho \in \mathbb{H}$ be such that $\rho^n = 1$ for positive integer n. In the spirit of Cullen [7] write $\rho = \rho_0 + \rho_I I$ where $\rho_I = \sqrt{\rho_1^2 + \rho_2^2 + \rho_3^2}$ and $I = \rho_I^{-1}(\rho_1 \mathbf{i} + \rho_2 \mathbf{j} + \rho_3 \mathbf{k})$ satisfies $I^2 = -1$. Since ρ is assumed to be in the complex plane where the entire iteration occurs, I defines the imaginary direction of that plane. We now rewrite ρ as $\rho = e^{I\theta}$ and demand that $e^{In\theta} = 1$, with the usual result that $\theta = 2\pi m/n$ for integer m. In terms of ρ_0 and ρ_I we have $\rho_0 = \cos(2\pi m/n)$ and $\rho_I = \sin(2\pi m/n)$. When $n = 2$ this gives only the two usual square-roots of unity, ± 1, and the result is a single bifurcation point on the real axis (as with the complex Mandelset), however, for higher n we now have $\rho_1^2 + \rho_2^2 + \rho_3^2 = \sin^2(2\pi m/n)$, which results in a whole 3-sphere of nth roots, all with real part $\rho_0 = \cos(2\pi m/n)$. The implications of this fact for the Mandelset are portrayed in Figures 2–6 and described in the captions.

All n-furcation points along the real axis remain as points, but those offset from the real axis become 3-spheres of contact between regions of different periods.

The preceding analysis reveals that the 'Mandelset' for the function $q^2 + c$ has no special properties not already recognized in the complex case. In particular, from a bifurcation point of view, the imaginary directions $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are equivalent. Nevertheless, it remains an entertaining graphical exercise to portray different aspects of the set as well as of the Julia sets associated with the iteration **[12, 13, 14]**. Various pictures of such Julia sets are also available on the internet, and can be found by using the search facility of netbrowsers.

Many of the deeper mathematical theorems associated with the complex Mandelset rely heavily on the property of analyticity possessed by polynomial and transcendental functions of z alone. We have already seen that quaternion polynomials are not regular with respect to the naturally defined CRF operator (although polynomials constructed with sums of terms of the form $q^n a$, $(a \in \mathbb{H})$ are Cullen-regular). We should therefore next ask if we can find regular functions of q that iterate in a more interesting way. For Cullen-regular quadratics, we have demonstrated **[11]** that this definition of regularity results in functions that can be considered complex-equivalent from an iterative point of view. When written in polar notation, all such functions take the form $f(q) = \alpha(q_0, r) + \beta(q_0, r) r^{-1} \mathsf{Pu}(q)$, where $r = \sqrt{(q_1^2 + q_2^2 + q_3^2)}$ and α, β are functions determined by the form of f. In effect, the term $r^{-1}\mathsf{Pu}(q)$ plays the role of an imaginary unit and, along with the parameter q_0, serves to define a complex subspace of $\mathbb{H}$ within which all the iteration takes place. Once again, therefore, we find nothing specifically quaternionic in the iterative behavior.

If instead of the Cullen definition of regularity, we revert to the CRF version, then as already noted, polynomial functions of q (with real coefficients) are not regular. Whereas Mandelsets resulting from the iteration of such functions do appear interesting, it seems very unlikely that the theorems and techniques that would be applicable in the complex case can be made to apply for such functions in $\mathbb{H}$. In particular, results associated with normal forms cannot be reproduced for the quaternions. The most general regularly iterable *linear* maps have been exhibited **[10]** but have been shown to possess no interesting dynamical behavior. A study **[11]** of regularly iterable quadratic and higher power maps leads only to maps that are immediately restricted to a complex subspace.

We have restricted ourselves to an investigation of the iterative behavior of quaternion maps that satisfy various definitions of regularity and have found this constraint always to be too strong to obtain interesting dynamical behavior. At this point, one might decide to forgo the demand of any kind of regularity, but we

re-emphasize our observation that in the complex case all the fascinating dynamical behavior is intimately intertwined with the analyticity condition, which has the effect of staking a claim on ground intermediate between $\mathbb{R}$ and $\mathbb{R}^2$. On the other hand, functions on quaternion space are equivalent to functions on $\mathbb{R}^4$ and an insistence on regularity fails to find any intermediate ground between $\mathbb{R}^4$ and $\mathbb{R}^2$. For these reasons we see little value in further pursuing these studies.

ACKNOWLEDGMENTS. Most of this work was done while the second author was at the Department of Mathematics, University of Melbourne. Useful comments on an early draft of this article came from Ralph Heidrich of the Mathematisch-Naturwissenschaftliche Fakultät der Rheinisch-Westfälischen Technischen Hochschule, Aachen.

REFERENCES

1. A. Sudbery, Quaternionic Analysis, *Proc. Camb. Phil. Soc.* **85** (1979), 199–225.
2. M. Koecher and R. Remmert, Hamilton's quaternions, Chapter 7 in *Numbers*, edited by J. Ewing, Springer, New York 1990.
3. R. Fueter, Analytische Funktionen einer Quaternionenvariablen, *Comment. Math. Helv.* **4** (1932), 9–20.
4. R. Fueter, Die Funktionentheorie der Differentialgleichungen, *Comment. Math. Helv.* **7** (1934), 307–330.
5. R. Fueter, Über die analytische Darstellung der regulären Funktionen einer Quaternionenvariablen, *Comment. Math. Helv.* **8** (1935/6), 371–378.
6. R. F. Rinehart, Elements of a Theory of Intrinsic Functions on Algebras, *Duke Math. J.* **27** (1960), 1–19.
7. C. G. Cullen, An Integral Theorem for Analytic Intrinsic Functions on Quaternions, *Duke Math. J.* **32** (1965), 139–148.
8. R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley, second edition, Redwood City, 1989.
9. R. L. Devaney and L. Keen, editors, *Chaos and Fractals*: *The mathematics behind the computer graphics*, volume 39 of *Proceedings of Symposia in Applied Mathematics*, American Mathematical Society, Providence, 1989.
10. S. P. Bedding and K. M. Briggs, Regularly Iterable Linear Quaternion Maps, La Trobe University Mathematics Research Report No. 95-2, Bundoora, Victoria 3083, Australia, 1995.
11. S. P. Bedding and K. M. Briggs, Iteration of Quaternion Maps, *Int. J. Bifurcation and Chaos* **5** (1995), 877–881.
12. J. A. R. Holbrook, Quaternionic Fatou-Julia Sets, Les Annales des Sciences Mathématiques de Québec **11** (1987), 79–94.
13. A. Norton, Julia Sets in the Quaternions, *Computers and Graphics* **13** (1989), 267–278.
14. J. Kozak and P. Petek, On the iteration of a quadratic family in quaternions, Technical report, Department of Mathematics, University of Ljubljana, University of Ljubljana, Jadranska 19, 6111 Ljubljana, Slovenia, 1992.

Bedding:
School of Mathematics
La Trobe University
Bundoora, Victoria 3083
AUSTRALIA
s.bedding@latrobe.edu.au

Briggs:
Department of Physics
University of Western Australia
Nedlands, WA 6907
AUSTRALIA
keith.briggs@physics.uwa.edu.au

The Fifty-Sixth William Lowell Putnam Mathematical Competition

Leonard F. Klosinski, Gerald L. Alexanderson, and Loren C. Larson

The following are the results of the fifty-sixth William Lowell Putnam Mathematical Competition, held on December 2, 1995. They have been determined in accordance with the regulations governing the Competition, an annual contest supported by the William Lowell Putnam Prize Fund for the Promotion of Scholarship, a fund left by Mrs. Putnam in memory of her husband. The Competition is held under the auspices of the Mathematical Association of America.

The first prize, $7,500, was awarded to the Department of Mathematics at Harvard University. The members of the winning team were Kiran S. Kedlaya, Lenhard L. Ng, and Hong Zhou; each was awarded a prize of $500.

The second prize, $5,000, was awarded to the Department of Mathematics at Cornell University. The members of the winning team were Jeremy L. Bem, Robert D. Kleinberg, and Mark Krosky; each was awarded a prize of $400.

The third prize, $3,000, was awarded to the Department of Mathematics at the Massachusetts Institute of Technology. The members of the winning team were Ruth A. Britto-Pacumio, Sergey M. Ioffe, and Thomas A. Weston; each was awarded a prize of $300.

The fourth prize, $2,000, was awarded to the Department of Mathematics at the University of Toronto. The members of the winning team were Edward Goldstein, J. P. Grossman, and Naoki Sato; each was awarded a prize of $200.

The fifth prize, $1,000, was awarded to the Department of Mathematics at Princeton University. The members of the winning team were Michael J. Goldberg, Alex Heneveld, and Jacob A. Rasmussen; each was awarded a prize of $100.

The five highest ranking individual contestants, in alphabetical order, were Yevgeniy Dodis, New York University; J. P. Grossman, University of Toronto; Kiran S. Kedlaya, Harvard University; Sergey V. Levin, Harvard University; and Lenhard L. Ng, Harvard University. Each of these has been designated a Putnam Fellow by the Mathematical Association of America and awarded a prize of $1,000 by the Putnam Prize Fund.

The next four highest ranking contestants, in alphabetical order, were Mark Krosky, Cornell University; Serban M. Nacu, Harvard University; Akira Negi, University of North Carolina, Chapel Hill; and Chung-Chieh Shan, Harvard University; each was awarded a prize of $500.

The next five highest ranking contestants, in alphabetical order, were Aaron F. Archer, Harvey Mudd College; Jeremy L. Bem, Cornell University; Robert D. Kleinberg, Cornell University; David L. Savitt, University of British Columbia; and Hong Zhou, Harvard University; each was awarded a prize of $250.

The next ten highest ranking individuals, in alphabetical order, are Federico Ardila, Massachusetts Institute of Technology; Robert G. Au, Stanford University; Ioana Dumitriu, New York University; Craig R. Helfgott, Princeton University; John J. Krueger, Hope College; Daniel K. Schepler, Washington University, St. Louis; Mikhail V. Shubov, Texas Tech University; Balint Virag, Harvard University; Ronald A. Walker, University of Richmond; and Stephen S. Wang, Harvard University; each was awarded a prize of $100.

The following teams, named in alphabetical order, received honorable mention; University of British Columbia, with team members Mark Hamilton, Erich Mueller, and David L. Savitt; University of Chicago, with team members Dean Jens, Christopher D. Jeris, and Jeffrey Willson; Duke University, with team members Johanna Miller, Noam Shazeer, and Tung Tran; New York University, with team members Yevgeniy Dodis, Ioana Dumitriu, and Yevgeniy Kovchegov; and Washington University, St. Louis, with team members Matthew Crawford, Daniel K. Schepler, and Jade P. Vinson.

Honorable mention was achieved by the following thirty-two individuals named in alphabetical order: Jared E. Anderson, University of Victoria; Donald A. Barkauskas, Rice University; William J. Beckler, Colgate University; Manjul Bhargava, Harvard University; David E. Bradley, Virginia Polytechnic Institute and State University; Ruth A. Britto-Pacumio, Massachusetts Institute of Technology; Samit Dasgupta, Harvard University; Todd W. Geldon, Princeton University; Andrei C. Gnepp, Harvard University; Michael J. Goldberg, Princeton University; Wei-Hwa Huang, California Institute of Technology; Sergey M. Ioffe, Massachusetts Institute of Technology; David Y. Jao, Massachusetts Institute of Technology; Christopher D. Jeris, University of Chicago; Sergey Kirshner, University of California, Berkeley; Mikhail G. Konikov, University of Maryland, College Park; Eric H. Kuo, Massachusetts Institute of Technology; Frédéric Latour, University of Waterloo; Dion Lew, University of Toronto; Adam W. Meyerson, Massachusetts Institute of Technology; Olexei I. Motrunich, University of Missouri, Columbia; Roman Muchnik, California Institute of Technology; Colin A. Percival, Simon Fraser University; Rajesh J. Pereira, McGill University; Jacob A. Rasmussen, Princeton University; Naoki Sato, University of Toronto; Douglas Squirrel, Reed College; Guido Ubaldis, University of California, Berkeley; Jade P. Vinson, Washington University, St. Louis; Jonathan L. Weinstein, Harvard University; Thomas A. Weston, Massachusetts Institute of Technology; and Liang Yang, Yale University.

The other individuals who achieved ranks among the top 102, in alphabetical order of their schools, were: University of British Columbia, Erich J. Mueller; University of California, Santa Barbara, Akshay Venkatesh; Carleton University, Bhaskara M. Marthi; University of Connecticut, William M. Watson; University of Delaware, Charles W. Helms; Duke University, Johanna L. Miller; Harvard University, Matthew L. Bruce, Hank S. Chien, Patrick K. Corn, Adam Kalai, Paul Li, David L. McAdams, Daniel S. Quint, Robert Ribciuc, Scott R. Sheffield, Florin Spinu, Eric G. Yeh; University of Illinois, Champaign-Urbana, Tsz Ho Chan; Massachusetts Institute of Technology, Matthew D. Blum, Amit Khetan, Sergei Krupenin, Alexander Morcos; McGill University, François Labelle; University of Missouri, Rolla, Hal J. Burch; City University of New York, Queen's College, Daniil Khaykis; Princeton University, Peter A. Coles, Alex Heneveld, Michael Krasnitz, Andrew M. Neitzke, Alexandru-Anton A. M. Popa, Ransom L. Richardson; Queen's University, Joanna Karczmarek; Reed College, Galen B. Huntington; Rice University, Ron D. Dror, Brian M. Wahlert; Stanford University,

Christopher C. Cheng, Theodore H. Hwa; University of Toronto, Edward Goldstein, Cyrus Hsia, Alexander M. Nicholson; Vassar College, Andrew F. Rizzo; University of Victoria, Peter J. Dukes; University of Waterloo, Jeff W. Brown, Peter L. Milley, Wei Yu; and Whittier College, Joshua S. Worley.

The Elizabeth Lowell Putnam Prize, named for the wife of William Lowell Putnam and to be "awarded periodically to a woman whose performance on the Competition has been deemed particularly meritorious," is awarded this year to Ioana Dumitriu of New York University. The winner is awarded a prize of $500.

There were 2,468 individual contestants from the 405 colleges and universities in Canada and the United States in the competition of December 2, 1995. Teams were entered by 306 institutions. The Questions Committee for the fifty-sixth competition consisted of Fan Chung, University of Pennsylvania, chair; Mark I. Krusemeyer, Carleton College, and Richard K. Guy, University of Calgary; they composed the problems listed below and were most prominent among those suggesting solutions.

PROBLEMS

Problem A-1. Let S be a set of real numbers which is closed under multiplication (that is, if a and b are in S, then so is ab). Let T and U be disjoint subsets of S whose union is S. Given that the product of any *three* (not necessarily distinct) elements of T is in T and that the product of any three elements of U is in U, show that at least one of the two subsets T, U is closed under multiplication.

Problem A-2. For what pairs (a, b) of positive real numbers does the improper integral

$$\int_b^{\infty}\left(\sqrt{\sqrt{x+a}-\sqrt{x}}-\sqrt{\sqrt{x}-\sqrt{x-b}}\right)dx$$

converge?

Problem A-3. The number $d_1d_2\dots d_9$ has nine (not necessarily distinct) decimal digits. The number $e_1e_2\dots e_9$ is such that each of the nine 9-digit numbers formed by replacing just one of the digits d_i in $d_1d_2\dots d_9$ by the corresponding digit e_i $(1 \le i \le 9)$ is divisible by 7. The number $f_1f_2\dots f_9$ is related to $e_1e_2\dots e_9$ in the same way: that is, each of the nine numbers formed by replacing one of the e_i by the corresponding f_i is divisible by 7. Show that, for each i, $d_i - f_i$ is divisible by 7.

[For example, if $d_1d_2\dots d_9 = 199501996$, then e_6 may be 2 or 9, since 199502996 and 199509996 are multiples of 7.]

Problem A-4. Suppose we have a necklace of n beads. Each bead is labeled with an integer and the sum of all these labels is $n-1$. Prove that we can cut the necklace to form a string whose consecutive labels $x_1, x_2, \dots, x_n$ satisfy

$$\sum_{i=1}^{k} x_i \le k-1 \qquad \text{for} \quad k = 1, 2, \dots, n.$$

Problem A-5. Let $x_1, x_2, \ldots, x_n$ be differentiable (real-valued) functions of a single variable t which satisfy

$$\begin{aligned}
\frac{dx_1}{dt} &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\
\frac{dx_2}{dt} &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\
&\vdots \\
\frac{dx_n}{dt} &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n
\end{aligned}$$

for some constants $a_{ij} \geq 0$. Suppose that for all i, $x_i(t) \to 0$ as $t \to \infty$. Are the functions $x_1, x_2, \ldots, x_n$ necessarily linearly dependent?

Problem A-6. Suppose that each of n people writes down the numbers $1, 2, 3$ in random order in one column of a $3 \times n$ matrix, with all orders equally likely and with the orders for different columns independent of each other. Let the row sums a, b, c of the resulting matrix be rearranged (if necessary) so that $a \leq b \leq c$. Show that for some $n \geq 1995$, it is at least four times as likely that both $b = a + 1$ and $c = a + 2$ as that $a = b = c$.

Problem B-1. For a partition π of $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, let $\pi(x)$ be the number of elements in the part containing x. Prove that for any two partitions π and π', there are two distinct numbers x and y in $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ such that $\pi(x) = \pi(y)$ and $\pi'(x) = \pi'(y)$. [A *partition* of a set S is a collection of disjoint subsets (parts) whose union is S.]

Problem B-2. An ellipse, whose semi-axes have lengths a and b, rolls without slipping on the curve $y = c\sin(x/a)$. How are a, b, c related, given that the ellipse completes one revolution when it traverses one period of the curve?

Problem B-3. To each positive integer with n^2 decimal digits we associate the determinant of the matrix obtained by writing the digits in order across the rows. For example, for $n = 2$, to the integer 8617 we associate $\det\begin{pmatrix} 8 & 6 \\ 1 & 7 \end{pmatrix} = 50$. Find, as a function of n, the sum of all the determinants associated with n^2-digit integers. (Leading digits are assumed to be nonzero; for example, for $n = 2$, there are 9000 determinants.)

Problem B-4. Evaluate

$$\sqrt[8]{2207 - \cfrac{1}{2207 - \cfrac{1}{2207 - \cdots}}}.$$

Express your answer in the form $(a + b\sqrt{c})/d$, where a, b, c, d are integers.

Problem B-5. A game starts with four heaps of beans, containing 3, 4, 5 and 6 beans. The two players move alternately. A move consists of taking *either*

a. one bean from a heap, provided at least two beans are left behind in that heap, or
b. a complete heap of two or three beans.

The player who takes the last heap wins. To win the game, do you want to move first or second? Give a winning strategy.

Problem B-6. For a positive real number α, define

$$S(\alpha) = \{\lfloor n\alpha \rfloor : n = 1,2,3,\dots\}.$$

Prove that $\{1,2,3,\dots\}$ cannot be expressed as the disjoint union of three sets $S(\alpha)$, $S(\beta)$, and $S(\gamma)$. [As usual, $\lfloor x \rfloor$ is the greatest integer $\leq x$.]

SOLUTIONS. In the 12-tuples $(n_{10}, n_9, n_8, n_7, n_6, n_5, n_4, n_3, n_2, n_1, n_0, n_{-1})$ following each problem number below, n_i for $10 \geq i \geq 0$ is the number of students among the top 204 contestants achieving i points for the problem and n_{-1} is the number of those not submitting solutions.

A-1 (115, 64, 20, 0, 0, 0, 0, 0, 1, 0, 2, 2)

Solution. Suppose T is not closed under multiplication. Then there are elements $t_1, t_2 \in T$ with $t_1 t_2 \notin T$, and since S is closed under multiplication, $t_1 t_2 \in U$. Now consider any two elements $u_1, u_2 \in U$; we'll show that $u_1 u_2 \in U$ (and thus that U is closed under multiplication). Suppose $u_1 u_2 \notin U$. Then $u_1 u_2 \in T$, so $t_1 \cdot t_2 \cdot u_1 u_2 \in T$ (as a product of three elements of T), but also $t_1 \cdot t_2 \cdot u_1 u_2 = (t_1 t_2) \cdot u_1 \cdot u_2 \in U$ (as a product of three elements of U), which is a contradiction since T and U are disjoint. So $u_1 u_2 \in U$, and we are done.

Comment. T, U need not both be closed; for example, if T (respectively U) is the set of positive integers $\equiv 1$ (respectively 3) modulo 4, then $S = T \cup U$ and T are closed, but U is not.

A-2 (26, 10, 14, 0, 0, 0, 0, 0, 14, 3, 56, 81)

Solution 1. The integral converges if and only if $a = b$.

Note that the integrand is defined and continuous for all $x \geq b$, so the only issue is convergence at ∞.

Since $(\sqrt{x+a} - \sqrt{x})(\sqrt{x+a} + \sqrt{x}) = a$ and $(\sqrt{x} - \sqrt{x-b})(\sqrt{x} + \sqrt{x-b}) = b$, we can rewrite the integrand as

$$\frac{\sqrt{a}}{\sqrt{\sqrt{x+a} + \sqrt{x}}} - \frac{\sqrt{b}}{\sqrt{\sqrt{x} + \sqrt{x-b}}}$$

$$= \frac{1}{\sqrt[4]{x}} \left[\frac{\sqrt{a}}{\sqrt{\sqrt{1 + \frac{a}{x}} + 1}} - \frac{\sqrt{b}}{\sqrt{1 + \sqrt{1 - \frac{b}{x}}}} \right].$$

If $a \neq b$, the quantity in the brackets approaches the nonzero number $\sqrt{a/2} - \sqrt{b/2}$ as $x \to \infty$, so for large enough x the absolute value of the integrand is at least $c/2\sqrt[4]{x}$, where $c = |\sqrt{a/2} - \sqrt{b/2}|$. The integrand then diverges by comparison with the divergent integral $\int_1^\infty (1/\sqrt[4]{x})\,dx$.

On the other hand, if $a = b$, the integrand equals

$$\frac{\sqrt{a}}{\sqrt[4]{x}}\left(\frac{1}{\sqrt{\sqrt{1+\frac{a}{x}}+1}}-\frac{1}{\sqrt{1+\sqrt{1-\frac{a}{x}}}}\right)$$

$$=\frac{\sqrt{a}}{\sqrt[4]{x}}\cdot\frac{1}{\sqrt{\sqrt{1+\frac{a}{x}}+1}\sqrt{1+\sqrt{1-\frac{a}{x}}}}\left(\sqrt{1+\sqrt{1-\frac{a}{x}}}-\sqrt{\sqrt{1+\frac{a}{x}}+1}\right)$$

$$=\frac{\sqrt{a}}{\sqrt[4]{x}}\cdot\frac{1}{\sqrt{\sqrt{1+\frac{a}{x}}+1}\sqrt{1+\sqrt{1-\frac{a}{x}}}}\,\frac{1+\sqrt{1-\frac{a}{x}}-\left(\sqrt{1+\frac{a}{x}}+1\right)}{\sqrt{1+\sqrt{1-\frac{a}{x}}}+\sqrt{\sqrt{1+\frac{a}{x}}+1}}$$

$$=\frac{\sqrt{a}}{\sqrt[4]{x}}\cdot\frac{1}{\varphi(x)}\left(\sqrt{1-\frac{a}{x}}-\sqrt{1+\frac{a}{x}}\right)$$

$$=\frac{\sqrt{a}}{\sqrt[4]{x}}\cdot\frac{1}{\varphi(x)}\cdot\frac{1-\frac{a}{x}-\left(1+\frac{a}{x}\right)}{\sqrt{1-\frac{a}{x}}+\sqrt{1+\frac{a}{x}}}=\frac{-2a\sqrt{a}}{x\sqrt[4]{x}\,\varphi(x)\left(\sqrt{1-\frac{a}{x}}+\sqrt{1+\frac{a}{x}}\right)},$$

where

$$\varphi(x)=\sqrt{\sqrt{1+\frac{a}{x}}+1}\sqrt{1+\sqrt{1-\frac{a}{x}}}\left(\sqrt{1+\sqrt{1-\frac{a}{x}}}+\sqrt{\sqrt{1+\frac{a}{x}}+1}\right).$$

Since $\varphi(x)\to\sqrt{2}\cdot\sqrt{2}\cdot(\sqrt{2}+\sqrt{2})=4\sqrt{2}$ as $x\to\infty$, the absolute value of the integrand is then less than $a\sqrt{a}/x^{5/4}$, and the integral converges by comparison with $\int_1^\infty(1/x^{5/4})\,dx$.

Solution 2. The integrand equals

$$x^{1/4}\left(\sqrt{\left(1+\frac{a}{x}\right)^{1/2}-1}-\sqrt{1-\left(1-\frac{b}{x}\right)^{1/2}}\right)$$

$$=x^{1/4}\left(\sqrt{\frac{a}{2x}-\frac{a^2}{8x^2}+O(x^{-3})}-\sqrt{\frac{b}{2x}+\frac{b^2}{8x^2}+O(x^{-3})}\right)$$

$$=x^{-1/4}\left(\sqrt{\frac{a}{2}}\sqrt{1-\frac{a}{4x}+O(x^{-2})}-\sqrt{\frac{b}{2}}\sqrt{1+\frac{b}{4x}+O(x^{-2})}\right)$$

$$=x^{-1/4}\left(\sqrt{\frac{a}{2}}\left[1-\frac{a}{8x}+O(x^{-2})\right]-\sqrt{\frac{b}{2}}\left[1+\frac{b}{8x}+O(x^{-2})\right]\right)$$

$$=x^{-1/4}\left(\sqrt{\frac{a}{2}}-\sqrt{\frac{b}{2}}\right)+O(x^{-5/4}),$$

so since $\int_1^\infty x^{-1/4}\,dx$ diverges and $\int_1^\infty x^{-5/4}\,dx$ converges, the integral converges just if $a = b$.

Solution 3. Let $f(x) = \sqrt{x}$. By the Mean Value Theorem there is a number h (which depends on x), $0 < h < a$, such that $f(x+a) - f(x) = f'(x+h)a$, and a number k (which depends on x), $0 < k < b$, such that $f(x) - f(x-b) = f'(x-k)b$. Thus,

$$\sqrt{x+a} - \sqrt{x} = \frac{a}{2\sqrt{x+h}} \qquad \text{for some } h, 0 < h < a, \quad \text{and}$$

$$\sqrt{x} - \sqrt{x-b} = \frac{b}{2\sqrt{x-k}} \qquad \text{for some } k, 0 < k < b.$$

It follows that

$$\sqrt{\sqrt{x+a} - \sqrt{x}} - \sqrt{\sqrt{x} - \sqrt{x-b}}$$

$$= \frac{\sqrt{a}}{\sqrt{2}\,(x+h)^{1/4}} - \frac{\sqrt{b}}{\sqrt{2}\,(x-k)^{1/4}}$$

$$= \frac{(x-k)^{1/4}\sqrt{a} - (x+h)^{1/4}\sqrt{b}}{\sqrt{2}\,(x+h)^{1/4}(x-k)^{1/4}}$$

$$= \frac{(a^2-b^2)x - (ka^2+hb^2)}{\sqrt{2}\,(x+h)^{1/4}(x-k)^{1/4}\left[(x-k)^{1/4}\sqrt{a} + (x+h)^{1/4}\sqrt{b}\right]\left[(x-k)^{1/2}a + (x+h)^{1/2}b\right]}.$$

Thus, the integrand is $O(x^{-1/4})$ when $a \neq b$ and $O(x^{-5/4})$ when $a = b$. The result follows as in the previous solutions.

A-3 (95, 44, 39, 0, 0, 0, 0, 0, 12, 5, 3, 6)

Solution. Suppose $d_1 d_2 \ldots d_9 \equiv a \pmod 7$. Then for each i, $1 \leq i \leq 9$,

$$\begin{aligned} a \equiv a - 0 &\equiv (d_1 \ldots d_i \ldots d_9) - (d_1 \ldots e_i \ldots d_9) \pmod 7 \\ &= 10^{9-i}d_i - 10^{9-i}e_i \pmod 7. \end{aligned}$$

On summing these congruences, we find that $9a \equiv (d_1 d_2 \ldots d_9) - (e_1 e_2 \ldots e_9) \pmod 7$, and therefore $e_1 e_2 \ldots e_9 \equiv -a \pmod 7$.

In a similar manner, starting with $e_1 e_2 \ldots e_9 \equiv -a \pmod 7$, we have

$$-a \equiv 10^{9-i}e_i - 10^{9-i}f_i \pmod 7,$$

and therefore

$$-a \equiv \left(10^{9-i}d_i - a\right) - 10^{9-i}f_i \pmod 7,$$

or equivalently,

$$10^{9-i}d_i \equiv 10^{9-i}f_i \pmod 7.$$

Since 7 and 10 are relatively prime, $d_i \equiv f_i \pmod 7$.

A-4 (39, 9, 13, 0, 0, 0, 0, 0, 19, 1, 83, 40)

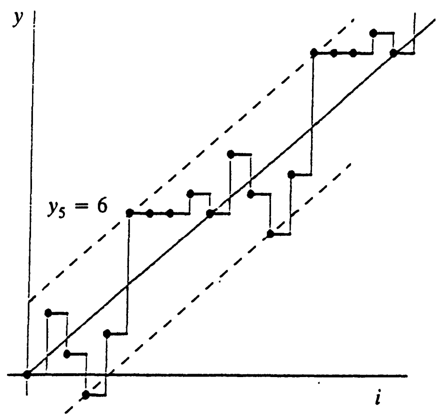
Solution 1. We will show that there are just two places where we may cut the necklace. Each is associated with the sense in which we go around the necklace.

Choose an arbitrary starting position, and a sense of rotation, and let the labels be the integers $y_1, y_2, \ldots, y_n$, where $\sum_{i=1}^{n} y_i = n - 1$.

Consider the path in the coordinate plane which starts from the origin, $(0,0)$, moves one space to the right and then vertically to the point $(1, y_1)$, then one space

to the right and vertically to the point $(2, y_1 + y_2)$, and so on to the points $(3, y_1 + y_2 + y_3), \ldots, (k, \Sigma_{i=1}^k y_i), \ldots, (n, \Sigma_{i=1}^n y_i = n - 1)$. Continue on around the necklace, repeating the pattern: $(n + 1, n - 1 + y_1), (n + 2, n - 1 + y_1 + y_2), \ldots$. Choose the point(s) $(K, \Sigma_{i=1}^K y_i)$ which maximize the height above the line $y = ((n - 1)/n)x$; that is, maximize $\Sigma_{i=1}^k y_i - ((n - 1)/n)k$. Since $n - 1$ and n are relatively prime integers, K is unique modulo n. Relabel the integers, decreasing the subscripts by K, that is, move the origin to the chosen point. Since the slope of the line from this point to any other is at most $(n - 1)/n$ (with equality only at the end of each period), we have achieved our aim: if $\Sigma_{i=1}^k x_i \geq k$, the slope would be at least 1.

Example. Suppose $\{y_1, y_2, \ldots, y_n\} = \{3, -2, -2, 3, 6, 0, 0, 1, -1\}$. The corresponding path is shown below.



The shifted sequence $0, 0, 1, -1, 3, -2, -2, 3, 6$ has partial sums $0, 0, 1, 0, 3, 1, -1, 2, 8 \leq k - 1 = 0, 1, 2, 3, 4, 5, 6, 7, 8$.

To find the cutting place associated with the opposite sense, we don't need to redraw the graph: simply select the *lowest* point below the line to get the shifted sequence $-2, -2, 3, -1, 1, 0, 0, 6, 3$ with partial sums $-2, -4, -1, -2, -1, -1, -1, 5, 8 \leq 0, 1, 2, 3, 4, 5, 6, 7, 8$.

Solution 2. Let the labels be denoted by $y_1, y_2, \ldots, y_n$, and suppose that there is no such k satisfying the above statement. Then for every k, there is a least number $\alpha_k \geq 1$ such that

$$\sum_{i=k}^{k-1+\alpha_k} y_i \geq \alpha_k,$$

where the subscripts i are taken modulo n. (Note that there are α_k terms in the sum.) We may assume that the labels are given so that $\alpha_1 \geq \alpha_i$ for $i = 1, 2, 3, \ldots, n$. We now choose numbers k_i as follows.

First, define $k_1 = 1$ and let β_1 be the smallest integer such that

$$\sum_{i=1}^{\beta_1} y_i \geq \beta_1.$$

(In this case, $\beta_1 = \alpha_1$.) Clearly $\beta_1 < n$. Thus, $I_1 = \{1, 2, \ldots, \beta_1\}$ is a proper subset of $\{1, 2, \ldots, n\}$.

Suppose that we have defined k_j, β_j, and I_j for $j < i$, with $k_1 < k_2 < \cdots < k_{i-1}$, and suppose we know that the I_j are disjoint sets whose union is $\{1, 2, \ldots, k_{i-1} - 1 + \beta_{i-1}\}$ and that this union is properly contained in $\{1, 2, \ldots, n\}$. Proceed as follows.

Define $k_i = k_{i-1} + \beta_{i-1}$, and choose β_i to be the smallest integer such that

$$\sum_{j=k_i}^{k_i-1+\beta_i} y_j \geq \beta_i.$$

(Note that $\beta_i = \alpha_{k_i}$.) Let $I_i = \{k_i, k_i + 1, \ldots, k_i - 1 + \beta_i\}$, and observe that I_i has β_i elements. We claim that $k_i - 1 + \beta_i < n$.

Case 1. Suppose that $k_i - 1 + \beta_i = n$. In this case,

$$\begin{aligned} n - 1 = \sum_{j=1}^{n} y_j &= \sum_{j=1}^{i-1} \left(\sum_{r=k_j}^{k_{j+1}-1} y_r \right) + \sum_{j=k_i}^{k_i-1+\beta_i} y_j \geq \sum_{j=1}^{i-1} \beta_j + \beta_i \\ &= \sum_{j=1}^{i-1} (k_{j+1} - k_j) + \beta_i = n, \end{aligned}$$

a contradiction.

Case 2. Suppose that $k_i - 1 + \beta_i > n$. Let $s = n + 1 - k_i$ and $t = k_i - 1 + \beta_i - n$ (note that $\beta_i = s + t$) and suppose that $t < \beta_1$. Then

$$\beta_i \leq \sum_{j=k_i}^{k_i-1+\beta_i} y_j = \sum_{j=k_i}^{k_i-1+s} y_j + \sum_{j=1}^{t} y_j \leq (s-1) + (t-1) = \beta_i - 2,$$

which is a contradiction. Therefore $t \geq \beta_1$ so that $\beta_i > \beta_1$. But this is a contradiction because $\beta_i = \alpha_{k_i} \leq \alpha_1 = \beta_1$.

We conclude that $k_i - 1 + \beta_i < n$, and so the I_i can never cover the entire set $\{1, 2, \ldots, n\}$, which is clearly absurd.

Therefore the necklace can be cut in the desired manner.

A-5 (1, 1, 2, 1, 0, 0, 0, 6, 18, 14, 43, 118)

Solution. We will show that the functions $x_1, x_2, \ldots, x_n$ are necessarily linearly dependent. Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}.$$

Then we are given that $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$. Consider a linear combination $y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ of the given functions, where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are constants, possibly complex, to be chosen later. If we set $\mathbf{v} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, we have $y = \mathbf{v} \cdot \mathbf{x} = \mathbf{v}^T \mathbf{x}$ (T = transpose) and thus

$$\frac{dy}{dt} = \mathbf{v}^T \frac{d\mathbf{x}}{dt} = \mathbf{v}^T \mathbf{A} \mathbf{x} = (\mathbf{A}^T \mathbf{v})^T \mathbf{x}.$$

In particular, if $\mathbf{v}$ is an eigenvector of $\mathbf{A}^T$ for the eigenvalue λ, we get

$$\frac{dy}{dt} = (\mathbf{A}^T \mathbf{v})^T \mathbf{x} = (\lambda \mathbf{v})^T \mathbf{x} = \lambda \mathbf{v} \cdot \mathbf{x} = \lambda y,$$

so in that case y has the form $y = Ce^{\lambda t}$ for some constant C.

Since we are given that $a_{ij} \geq 0$, in particular we have $\operatorname{Trace}(\mathbf{A}^T) = a_{11} + a_{22} + \cdots + a_{nn} \geq 0$, so $\mathbf{A}^T$ has at least one eigenvalue whose real part is nonnegative. Let λ be such an eigenvalue, and let $\mathbf{v} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ be a corresponding eigenvector of $\mathbf{A}^T$. Then, by the above, we have $y = Ce^{\lambda t}$ with $\operatorname{Re}(\lambda) \geq 0$. On the other hand, since $x_i(t) \to 0$ as $t \to \infty$ and y is a linear combination of the x_i, we have $y(t) \to 0$ as $t \to \infty$. But $|e^{\lambda t}| = e^{\operatorname{Re}(\lambda)t} \geq 1$ for $t \geq 0$, so $Ce^{\lambda t} \to 0$ implies $C = 0$. Therefore, a nontrivial linear combination of the x_i is identically zero (note that $\mathbf{v} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is nonzero because it is an eigenvector), so the x_i are linearly dependent, and we are done.

A-6 (1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 61, 141)

Solution. For a positive integer n, and using the other notation of the problem, let $S(n)$ be the statement "it is at least four times as likely that both $b = a + 1$ and $c = a + 2$ as that $a = b = c$." We show that if $S(n)$ is false, then $S(n + 1)$ is true. In particular, $S(n)$ is true for at least one of $n = 1995$, $n = 1996$.

For any positive integer n, let X_n be the number of ways the matrix can be formed so that $b = a + 1$ and $c = a + 2$ (where a, b, c, with $a \leq b \leq c$, are the row sums after rearrangement; incidentally, $a + b + c = 6n$). Let Y_n be the number of ways the matrix can be formed so that $a = b = c$, and let Z_n be the number of ways with $a = b$ and $c = a + 3$.

Our assumption that $S(n)$ is false means that $4Y_n > X_n$. Now note that if a matrix *with* $n + 1$ *columns* is formed such that its row sums are all equal, then the first n columns of that matrix form one of the matrices that is counted by X_n. Conversely, for each of the matrices counted by X_n, there is exactly one way to "complete" it to a matrix counted by Y_{n+1}, so we have $Y_{n+1} = X_n$. Similar arguments show that $Z_{n+1} \geq X_n$ (since to row sums $a, a + 1, a + 2$ one can add $2, 1, 3$ respectively to get $a + 2, a + 2, a + 5$), and $X_{n+1} \geq 6Y_n + 2X_n + 2Z_n$ (since $a + 2, a + 3, a + 4$ can be obtained by adding $1, 2, 3$ *in any order* to $a + 1, a + 1, a + 1$ [and rearranging], or by adding $2, 3, 1$ or $3, 1, 2$ (in that order) to $a, a + 1, a + 2$ [and rearranging], or by adding $3, 2, 1$ or $2, 3, 1$ to $a, a, a + 3$ [and rearranging]).

Therefore, we have

$$\begin{aligned}
\frac{X_{n+1}}{Y_{n+1}} = \frac{X_{n+1}}{X_n} &\geq 6\frac{Y_n}{X_n} + 2 + 2\frac{Z_n}{X_n} \\
&\geq 6\frac{Y_n}{X_n} + 2 + 2\frac{X_{n-1}}{X_n} = 6\frac{Y_n}{X_n} + 2 + 2\frac{Y_n}{X_n} \\
&= 8\frac{Y_n}{X_n} + 2.
\end{aligned}$$

But by our assumption, we have $Y_n/X_n > 1/4$, so $X_{n+1}/Y_{n+1} \geq 8/4 + 2 = 4$, so $X_{n+1} \geq 4Y_{n+1}$, and we are done.

B-1 (124, 26, 7, 0, 0, 0, 0, 0, 4, 10, 11, 22)

Solution. Suppose there are no two such numbers in $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Then any two numbers in the same part of π must be contained in parts of different size in π', and vice versa. This implies that the largest parts of π and π' have at most three numbers in them (because $1 + 2 + 3 + 4 > 9$). In fact, any two numbers in parts of the same size in π must be contained in parts of different sizes in π'. Therefore, π can have at most one part of size 3, one part of size 2, and at most three parts of size 1. This is impossible for a partition of a set of 9 numbers.

B-2 (2, 54, 26, 0, 0, 0, 0, 0, 3, 13, 51, 55)

Solution. We shall show that $b^2 = a^2 + c^2$.

The ellipse is given parametrically by the equations $x = a\cos\theta$, $y = b\sin\theta$. Using the familiar formula for arc length, the perimeter of the ellipse is

$$\int_0^{2\pi} \sqrt{a^2 \sin^2\theta + b^2\cos^2\theta}\, d\theta.$$

The length of one period of the sine curve is

$$\int_0^{2\pi a} \sqrt{1 + \frac{c^2}{a^2}\cos^2\left(\frac{x}{a}\right)}\, dx = \int_0^{2\pi} \sqrt{a^2 + c^2\cos^2\theta}\, d\theta.$$

Write $a^2 + c^2\cos^2\theta = a^2\sin^2\theta + (a^2 + c^2)\cos^2\theta$, and we see that the arc lengths will be equal if and only if $b^2 = a^2 + c^2$, as claimed.

B-3 (54, 15, 11, 0, 0, 0, 0, 0, 33, 15, 49, 27)

Solution. The sum is 45 for $n = 1$, 20250 for $n = 2$, and 0 for $n \geq 3$.

The case $n = 1$ is trivial: $1 + 2 + \cdots + 9 = 45$. Now let $n \geq 2$. Then for each $n \times n$ matrix with entries in $\{0, 1, 2, \ldots, 9\}$ there is another such matrix obtained by interchanging the last two columns. (If this matrix is equal to the original one, its determinant is zero.) Since interchanging two columns in a matrix changes its determinant to the opposite determinant, the sum of *all* determinants of matrices with entries in $\{0, 1, 2, \ldots, 9\}$ is zero. However, we are not supposed to take all such matrices, but only the ones that don't have a 0 in the upper left corner. If $n \geq 3$, interchanging the last two columns doesn't affect that corner, and so the required sum is 0 by the same argument. On the other hand, if $n = 2$, all determinants in the sum cancel *except* those of the form $\det\begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$. These determinants only depend on the diagonal entries, and there are ten of them for each pair of diagonal entries; thus their sum is

$$10 \sum_{i,j=1}^{9} ij = 10\left(\sum_{i=1}^{9} i\right)\left(\sum_{j=1}^{9} j\right) = 10 \cdot 45 \cdot 45 = 20250,$$

as claimed.

B-4 (9, 18, 62, 4, 0, 0, 0, 2, 21, 33, 10, 45)

Solution. The answer is $(3 + \sqrt{5})/2$.

Let

$$F(n) = n - \cfrac{1}{n - \cfrac{1}{n - \cdots}},$$

so the problem asks for $\sqrt[8]{F(2207)}$. Note that $F(n) = n - 1/F(n)$, and solving this quadratic equation for $F(n)$ yields $F(n) = (n \pm \sqrt{n^2 - 4})/2$.

For $n > 2$ we have

$$\frac{n - \sqrt{n^2 - 4}}{2} = \frac{2}{n + \sqrt{n^2 - 4}} < \frac{2}{n} < 1 < F(n),$$

so we must have the plus sign: $F(n) = (n + \sqrt{n^2 - 4})/2$.

Now note that

$$(F(n))^2 = \frac{n^2 + 2n\sqrt{n^2-4} + n^2 - 4}{4} = \frac{n^2 - 2 + n\sqrt{n^2-4}}{2}$$
$$= \frac{n^2 - 2 + \sqrt{(n^2-2)^2 - 4}}{2} = F(n^2 - 2)$$

for $n > 2$, since $n > 2$ implies $n^2 - 2 > 2$.

Conversely, if $k > 2$, then we have $k = n^2 - 2$ with $n = \sqrt{k+2} > 2$, and therefore $F(k) = (F(n))^2$, $\sqrt{F(k)} = F(n) = F(\sqrt{k+2})$. In particular,

$$\sqrt{F(2207)} = F(\sqrt{2209}) = F(47),$$
$$\sqrt[4]{F(2207)} = \sqrt{F(47)} = F(\sqrt{49}) = F(7), \quad \text{and}$$
$$\sqrt[8]{F(2207)} = \sqrt{F(7)} = F(\sqrt{9}) = F(3) = \frac{3+\sqrt{5}}{2}.$$

B-5 (72, 11, 13, 0, 0, 0, 0, 0, 9, 6, 42, 51)

Solution. Heaps of 0 or 1 cannot affect the game. In fact, heaps of 1 cannot arise. Heaps of 2 behave as though they were a single bean which may be removed. Heaps of 3 are special. Otherwise a move just removes a bean, and the result depends only on the parity of the total number of beans (counting a heap of 2 as a single bean).

The first player wins by taking one bean from the 3-heap, leaving heaps of 2, 4, 5 and 6 beans, whose "sum" is $1(=2) + 4 + 5 + 6$ which is even. Now the win is automatic, since the opponent must make the "sum" odd. It doesn't matter what moves are made, *except* that the first player mustn't move in a 4-heap (there is no need to since the sum will always be odd, and all the heaps can't be 4-heaps), and whenever the second player moves in a 4-heap, the first player removes all the remaining beans at the next move.

B-6 (3, 2, 0, 0, 0, 0, 0, 0, 2, 3, 61, 133)

Solution. Suppose $\alpha < \beta < \gamma$ and $S(\alpha)$, $S(\beta)$, and $S(\gamma)$ disjointly cover $\{1, 2, 3, \dots\}$. Since $\lfloor \alpha \rfloor = 1$, we have $\alpha = 1 + \epsilon$, for some ϵ satisfying $0 \le \epsilon < 1$.

Let $r > 1$ be the first value not in $S(\alpha)$. We have

$$\lfloor (r-1)\alpha \rfloor = r - 1, \quad \lfloor r\alpha \rfloor = r + 1.$$

Therefore,

$$(r-1)\alpha < r, \quad r\alpha \ge r + 1$$

and

$$1 + \frac{1}{r} \le \alpha < 1 + \frac{1}{r-1};$$

that is,

$$\frac{1}{r} \le \epsilon < \frac{1}{r-1}.$$

Fact 1. *If $u \notin S(\alpha)$, then the next element missing from $S(\alpha)$ is either $u + r$ or $u + r + 1$ (and the other of $u + r, u + r + 1$ is in $S(\alpha)$).*

Proof: Suppose $\lfloor t\alpha \rfloor = u - 1$, $\lfloor (t+1)\alpha \rfloor = u + 1$. Let $\delta = (t+1)\alpha - (u+1)$. The next missing element occurs at $u + m$ where m is the smallest integer such that $\delta + (m-1)\epsilon \geq 1$. If $m \leq r - 1$, we have

$$\delta + (m-1)\epsilon < m\epsilon \leq (r-1)\epsilon < 1$$

since $\delta < \epsilon$. Also, for $m = r + 1$,

$$\delta + (m-1)\epsilon = \delta + r\epsilon \geq 1.$$

Therefore $m = r$ or $r + 1$.

Note that $\lfloor \beta \rfloor = r$, so we have $r \leq \beta < r + 1$.

Fact 2. *If $v \in S(\beta)$, then the next element in $S(\beta)$ is $v + r$ or $v + r + 1$.*

Fact 2 can be proved in the same manner as Fact 1.

Combining Facts 1 and 2 with the fact that $S(\alpha)$ and $S(\beta)$ are disjoint, we conclude that the union of $S(\alpha)$ and $S(\beta)$ is all of $\{1, 2, 3, \dots\}$. Therefore, $\{1, 2, 3, \dots\}$ cannot be expressed as the disjoint union of *three* sets $S(\alpha)$, $S(\beta)$ and $S(\gamma)$.

Klosinski / Alexanderson:
Department of Mathematics
Santa Clara University
Santa Clara, CA 95053

Larson:
Department of Mathematics
St. Olaf College
Northfield, MN 55057

$\lim_{x \to \infty}(x^n/e^x) = 0$ for each $n = 1, 2, \dots$.

Because $e^x > x$ for all $x > 0$, the function $f(x) \equiv x/e^x$ is bounded on $(0, \infty)$. Then $g_n(x) \equiv x^n/e^x = n^n(f(x/n))^n$ is bounded on $(0, \infty)$ for each $n = 1, 2, \dots$ and $0 \leq x^n/e^x = g_{n+1}(x)/x \to 0$ as $x \to \infty$.

Contributed by Joseph M. Ling, University of Calgary

NOTES

Edited by: **John Duncan**

Geometric Properties of the Gamma Function

Patrick Ahern and Walter Rudin

The fact that $\log\Gamma(x)$ is convex on the positive real axis is one of the crucial properties of Γ. Combined with the functional equation $\Gamma(x+1) = x\Gamma(x)$ and the normalization $\Gamma(1) = 1$, it uniquely characterizes the gamma function. This theorem, due to Bohr and Mollerup, is the basis of Artin's elegant treatment of $\Gamma(x)$ in [2].

We extend this convexity property into the complex plane, use this extension to obtain information about the argument of $\Gamma(z)$ on vertical lines, and describe some features of the conformal mappings induced by Γ'/Γ and by $\log\Gamma$. In spite of the apparently inexhaustible supply of formulas, identities, expansions, and integrals concerned with Γ, we have found no mention of such geometric properties in the literature.

For brevity, we use the notation

$$G(z) = \log\Gamma(z).$$

We let Ω_a denote the right half-plane consisting of all complex z having $\operatorname{Re} z > a$, and denote its closure by $\overline{\Omega}_a$.

Theorem 1. (A) *If* $x \geq \frac{1}{2}$ *then* $\operatorname{Re} G''(x+iy) > 0$ *for all real* y.

(B) *If* $x < \frac{1}{2}$ *then* $\operatorname{Re} G''(x+iy) < 0$ *for all sufficiently large* y.

Note that $\operatorname{Re} G''(x+iy) = (\partial^2/\partial x^2)\log|\Gamma(x+iy)|$. The theorem asserts therefore *that* $\log|\Gamma(x+iy)|$ *is a convex function of* x *in* $\Omega_{1/2}$, *but in no larger half-plane.*

Proof: We begin with (B) since its proof gave us a hint that $1/2$ might be the cut-off between (A) and (B).

Define $\psi(s) = \frac{1}{2} - s$ on $(0,1]$, $\psi(s+1) = \psi(s)$ for all real s. One way of writing Stirling's formula is [4; p. 151]

$$\log\Gamma(z) = \left(z - \tfrac{1}{2}\right)\log z - z + \tfrac{1}{2}\log 2\pi + \int_0^\infty \frac{\psi(s)}{s+z}\,dx \tag{1}$$

for all $z \neq 0$ that are not negative real numbers. Differentiate this twice, then perform an integration by parts. The result is

$$G''(z) = \frac{1}{z} + \frac{1}{2z^2} + 6\int_0^\infty \frac{\varphi(s)}{(s+z)^4}\,ds, \tag{2}$$

where $\varphi(s) = \int_0^s \psi(t)\,dt$. Since ψ has mean-value 0, φ is periodic, hence bounded.

In fact, $0 \le \varphi(s) \le 1/8$. It follows that

$$\operatorname{Re} G''(x+iy) \le \frac{2x^3+x^2+(2x-1)y^2}{2(x^2+y^2)^2} + \frac{3}{4}\int_0^\infty \frac{ds}{\left[(s+x)^2+y^2\right]^2}.$$

The last integral is $O(y^{-3})$, as $y \nearrow \infty$. When $2x-1 \neq 0$, the dominant term, for fixed x and large y, is thus $(x-\frac{1}{2})y^{-2}$, and this is negative when $x < 1/2$; (B) follows.

Next, we apply log to the identity [**2**; p. 24]

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)},$$

then differentiate twice, and obtain

$$G''(z) + G''(1-z) = \frac{\pi^2}{\sin^2(\pi z)}.$$

Since $\sin \pi(\frac{1}{2}+iy) = \cos(\pi iy) = \cosh(\pi y)$, it follows that

$$G''\left(\tfrac{1}{2}+iy\right) + G''\left(\tfrac{1}{2}-iy\right) = \frac{\pi^2}{\cosh^2(\pi y)}$$

or

$$2\operatorname{Re} G''\left(\tfrac{1}{2}+iy\right) = \frac{\pi^2}{\cosh^2(\pi y)} > 0. \tag{3}$$

Finally, (2) shows that G'' is bounded in Ω_δ, for every $\delta > 0$. Since bounded harmonic functions in a half-plane are the Poisson integrals of their boundary values, (A) follows from (3).

The following monotonicity property was needed in [**1**], but only when $b-a$ is an integer, and in that case a more direct proof exists.

Theorem 2. (i) *If* $\frac{1}{2} \le a < b$ *then*

$$\arg\frac{\Gamma(b+iy)}{\Gamma(a+iy)}$$

is an increasing function of y *on* $(-\infty,\infty)$.

(ii) *The same conclusion holds if* $0 < a < \frac{1}{2}$ *and* $b > 1-a$.

Proof: Let u and v be the real and imaginary parts of $G = \log\Gamma$, respectively. Then $v = \arg\Gamma$, the Cauchy-Riemann equations give $u_x = v_y$, and hence $v_{xy} = u_{xx} > 0$ in $\overline{\Omega}_{1/2}$, by Theorem 1. This means that $v_y(a+iy) < v_y(b+iy)$, or

$$\frac{\partial}{\partial y}\arg\Gamma(a+iy) < \frac{\partial}{\partial y}\arg\Gamma(b+iy).$$

This proves (i).

To deduce (ii), note that

$$\frac{\Gamma(b+iy)}{\Gamma(a+iy)} = \frac{\Gamma(b+iy)}{\Gamma(1-a+iy)} \cdot \frac{\Gamma(1-a+iy)}{\Gamma(a+iy)} \cdot \frac{\Gamma(a-iy)}{\Gamma(a-iy)}$$
$$= \frac{1}{|\Gamma(a+iy)|^2} \cdot \frac{\Gamma(b+iy)}{\Gamma(1-a+iy)} \cdot \frac{\pi}{\sin\pi(a-iy)}. \tag{4}$$

Since now $1 - a > 1/2$, (ii) follows from (i) and the fact that the argument of the last factor in (4) is

$$\arctan[\cot(\pi a)\tanh(\pi y)],$$

which is an increasing function of y when $0 < a < 1/2$.

The proofs of the next two theorems will use the following sufficiency criterion for univalence.

Suppose: (a) Π *is an open half-plane that does not contain the origin*, (b) f *is holomorphic in a convex region* Ω, *and* (c) $f'(z) \in \Pi$ *for every* $z \in \Omega$. *Then* f *is univalent in* Ω.

When $\Pi = \Omega_0$ this is proved on p. 47 of [3]. It is clear that the criterion is rotation-invariant.

Theorem 3. (i) Γ'/Γ *is univalent in* Ω_0, *but in no larger half-plane.*

(ii) $\operatorname{Re}(\Gamma'/\Gamma)(x+iy) \geq (\Gamma'/\Gamma)(x)$ *in* Ω_0.

(iii) $\operatorname{Im}(\Gamma'/\Gamma)$ *is bounded in* Ω_δ, *for every* $\delta > 0$.

(iv) $|\operatorname{Im}(\Gamma'/\Gamma)(z)| < \pi/2$ *in* $\overline{\Omega}_{1/2}$.

Proof: Since $(\Gamma'/\Gamma)' = G''$, it follows from Theorem 1 and the sufficiency criterion that Γ'/Γ is univalent in $\Omega_{1/2}$. This proves only part of (i), but at least it points in the right direction.

Differentiation of $(\Gamma'/\Gamma)(x+iy) = (u+iv)(x+iy)$ (where u and v are now the real and imaginary parts of Γ'/Γ) with respect to y gives

$$iG''(x+iy) = (u_y + iv_y)(x+iy).$$

Thus $v_y = \operatorname{Re} G''$, hence (3) becomes

$$v_y\left(\tfrac{1}{2}+iy\right) = \frac{\pi^2}{2\cosh^2(\pi y)}.$$

If we integrate this we obtain

$$v\left(\tfrac{1}{2}+iy\right) = \frac{\pi}{2}\tanh(\pi y),$$

and this implies (iv).

Differentiation of the logarithm of

$$\frac{1}{\Gamma(z)} = ze^{\gamma z}\prod_1^\infty\left(1+\frac{z}{n}\right)e^{-z/n}$$

[4; p. 150], where γ is Euler's constant, leads to

$$u(x+iy) = -\gamma - \frac{x}{x^2+y^2} + \sum_1^\infty\left(\frac{1}{n} - \frac{n+x}{(n+x)^2+y^2}\right), \tag{5}$$
$$v(x+iy) = \sum_0^\infty \frac{y}{(n+x)^2+y^2}, \tag{6}$$

and

$$G''(z) = \sum_0^\infty \frac{1}{(n+z)^2}. \tag{7}$$

The right side of (5) is an increasing function of y^2, if $x > 0$. For fixed x it therefore attains its minimum when $y = 0$. This proves (ii), and (iii) follows from (6), because

$$v(x+iy) < \frac{y}{x^2+y^2} + \sum_1^\infty \frac{y}{n^2+y^2}$$
$$< \frac{1}{2x} + \int_0^\infty \frac{y}{t^2+y^2}\,dt = \frac{1}{2x} + \frac{\pi}{2}.$$

For the proof of (i), let $\Pi^+ = \{y > 0\}$ and $\Pi^- = \{y < 0\}$ be the upper and lower half-planes, respectively.

If $z \in \Omega_0 \cap \Pi^+$ then $(n+z)^2 \in \Pi^+$ for all $n \geq 0$, hence $(n+z)^{-2} \in \Pi^-$, so that $G''(z) \in \Pi^-$.

Likewise, $G''(z) \in \Pi^+$ if $z \in \Omega_0 \cap \Pi^-$.

It follows that Γ'/Γ is univalent in each of the quadrants $\Omega_0 \cap \Pi^+$ and $\Omega_0 \Pi^-$. By (6), Γ'/Γ maps $\Omega_0 \cap \Pi^+$ into Π^+ and $\Omega_0 \cap \Pi^-$ into Π^-; since Γ'/Γ is strictly increasing on the positive real axis, we conclude that Γ'/Γ is univalent in all of Ω_0.

Finally, Γ'/Γ has a pole at $z = 0$, hence maps each neighborhood of 0 onto a neighborhood of ∞. Since $(\Gamma'/\Gamma)(x) \nearrow \infty$ as $x \nearrow \infty$, it follows that Γ'/Γ is not univalent in Ω_δ if $\delta < 0$.

In the next theorem, x_0 is the unique positive number at which $\Gamma'(x_0) = 0$. Since $\Gamma(1) = \Gamma(2)$, $1 < x_0 < 2$.

Theorem 4. $\log \Gamma$ *is univalent in* Ω_{x_0}, *but in no larger half-plane.*

Proof: Recall that $(\log \Gamma)' = \Gamma'/\Gamma$. If $x > x_0$, it follows from part (ii) of Theorem 3 that

$$\operatorname{Re}(\log \Gamma)'(x+iy) \geq (\Gamma'/\Gamma)(x) > (\Gamma'/\Gamma)(x_0) = 0.$$

Thus $\log \Gamma$ is univalent in Ω_{x_0}. Clearly, x_0 has no neighborhood in which $\log \Gamma$ is univalent, since $(\log \Gamma)'(x_0) = 0$.

REFERENCES

1. P. Ahern, M. Flores, and W. Rudin, An invariant mean-value property, *J. Functional Anal.* **111**, 1993, 380–397.
2. E. Artin, *The Gamma Function*, Holt, Rinehart and Winston, 1964.
3. P. L. Duren, *Univalent Functions*, Springer-Verlag, 1983.
4. E. C. Titchmarsh, *The Theory of Functions*, 2nd ed., Oxford Univ. Press, 1939.
5. R. Remmert, Wielandt's theorem about the Γ-function, *Amer. Math. Monthly*, **103**, 1996, 214–220.

Department of Mathematics
University of Wisconsin-Madison
Madison, Wisconsin 53706-1388
ahern@math.wisc.edu

A Proof of the Existence of Infinite Product Probability Measures

Sadahiro Saeki

In memory of my dear friend Karl Stromberg

Let $\{(\Omega_i, \mathcal{F}_i, P_i): i \in I\}$ be a nonempty collection of probability spaces, and let $\Omega := \Pi_i \Omega_i$ be the product space. A measurable cylinder in Ω is a subset A of Ω of the form $A = \Pi_i A_i$, where $A_i \in \mathcal{F}_i$ for each i and $A_i = \Omega_i$ for all but finitely many i's. For such a set A, define $P(A) := \Pi_i P_i(A_i)$. By definition, the product probability measure of the P_i's is the (necessarily unique) extension of P to a probability measure on $\mathcal{F}(\mathcal{M}c)$, where $\mathcal{M}c$ is the collection of all measurable cylinders in Ω and $\mathcal{F}(\mathcal{M}c)$ is the σ-field generated by $\mathcal{M}c$. The standard proof of the existence of the product probability measure is based upon Fubini's Theorem for finitely many factors; see [**HS**: pp. 429–435]. We give a simple proof that does not require Fubini's Theorem.

Lemma. *Let $\mu: \mathcal{M}c \to [0,1]$ be a function such that $\sum_1^\infty \mu(A_n) = 1$ whenever (A_n) is a disjoint sequence in $\mathcal{M}c$ with union Ω. Then μ extends uniquely to a probability measure on $\mathcal{F}(\mathcal{M}c)$.*

Proof: Let $\mathcal{D}$ be the collection of all finite unions of measurable cylinders. It is easy to check that $\mathcal{D}$ is a field and each $A \in \mathcal{D}$ can be written as a finite disjoint union of members of $\mathcal{M}c$. In particular, A can be written as a countable disjoint union of members of $\mathcal{M}c$, say $A = \bigcup_1^\infty A_n$. Let $\mu'(A) := \sum_1^\infty \mu(A_n)$. To see that μ' is well-defined, write $\Omega \setminus A = \bigcup_1^m B_k$ with pairwise disjoint $B_k \in \mathcal{M}c$. Then

$$\sum_1^\infty \mu(A_n) = 1 - \sum_1^m \mu(B_k) \tag{1}$$

by our assumption on μ. Since the right-hand of (1) has nothing to do with the decomposition $\bigcup_1^\infty A_n$ of A, it follows that μ' is well-defined and therefore countably additive of $\mathcal{D}$. Hence the desired result is an immediate consequence of E. Hopf's extension theorem [**HS**: p. 142]. ∎

Theorem. *P extends uniquely to a probability measure on $\mathcal{F}(\mathcal{M}c)$.*

Proof: It suffices to prove that P satisfies the hypothesis of the lemma. Without loss of generality, assume that I is an infinite set. Let (A_n) be a disjoint sequence in $\mathcal{M}c$ with union Ω.

Case 1: I is countable. Then we may assume $I = \mathbb{N}$. Write $A_n = \prod_{i=1}^\infty A_{n,i}$, where $A_{n,i} \in \mathcal{F}_i$ for each i and $A_{n,i} = \Omega_i$ for all $i > i_n \in \mathbb{N}$. We claim that if $m \in \mathbb{N}$ and $x = (x_i)$ is an element of A_m and if $n \in \mathbb{N}$, then

$$\left\{ \prod_{i=1}^{i_m} \chi_{A_{n,i}}(x_i) \right\} \prod_{i > i_m} P_i(A_{n,i}) = \delta_{m,n} \quad \text{(Kronecker's delta)}. \tag{2}$$

For $n = m$, this is trivial, so assume $n \neq m$. Then, since $\Sigma_1^\infty \chi_{A_k} = 1$ identically and $\chi_{A_m}(x_1, \ldots, x_{i_m}, y_{i_m+1}, \ldots) = 1$ for all $y_i \in \Omega_i$ with $i > i_m$, we have

$$\left\{\prod_{i=1}^{i_m} \chi_{A_{n,i}}(x_i)\right\} \prod_{i>i_m} \chi_{A_{n,i}}(y_i) = 0 \tag{3}$$

for all such y_i's. Integrating each side of (3) finitely many times, we obtain (2) for $n \neq m$.

To get a contradiction, suppose $\Sigma_{n=1}^\infty P(A_n) \neq 1$. Then there must exist an $x_1 \in \Omega_1$ such that

$$\sum_{n=1}^\infty \chi_{A_{n,1}}(x_1) \prod_{i=2}^\infty P_i(A_{n,i}) \neq 1.$$

Hence an inductive argument yields an element $x = (x_i)$ of Ω such that

$$\sum_{n=1}^\infty \left\{\prod_{i=1}^{k} \chi_{A_{n,i}}(x_i)\right\} \prod_{i=k+1}^\infty P_i(A_{n,i}) \neq 1 \tag{4}$$

for each $k \geq 1$. But $x \in A_m$ for some $m \in \mathbb{N}$. Hence (4) with $k = i_m$ contradicts (2).

Case 2: I is uncountable. Then we can choose a countable subset J of I such that $A_n = A'_n \times \Omega'$ for all $n \geq 1$, where each A'_n is a measurable cylinder in $\Pi_{i \in J}\Omega_i$ and $\Omega' = \prod_{i \notin J}\Omega_i$. By Case 1 applied to (A'_n), we obtain $\Sigma_1^\infty P(A_n) = 1$. ■

Dedication. *Professor Karl Stromberg, my friend and colleague, died on July* 3, 1994. *He was an enthusiastic lover of the* Monthly. *When I presented the above proof in my seminar five to eight years ago, he liked it very much. Karl, I dedicate the present paper to you in the memory of our friendship. Have a peaceful sleep!*

REFERENCE

[HS] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, Berlin, 1965.

Department of Mathematics
Kansas State University
138 Cardwell Hall
Manhattan, KS 66506

A Problem

Leo S. Gurin

TRIBUTE. I learned about this problem and its solution in 1935, when I was in the eighth grade, from my teacher of mathematics, Yakov Stepanovich Chaikovsky, a very young man at that time. Now, in retrospective of a few decades of my own

professional work in mathematics, I understand that he was extremely talented. Unfortunately, he did not have a chance to develop his talent. During Stalin's terror in the later 1930s he was arrested and perished in the GULag. If he had a different fate, I am sure, his name would be listed among leading Soviet mathematicians. This note commemorates the 60th anniversary of the event when a young talented teacher demonstrated an elementary solution of an interesting problem to his students in a small village school.

PROBLEM. You have to expand functions $\sin x$ and $\cos x$ into series using only two facts from calculus:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

and

$$\lim_{m \to \infty} \frac{x^m}{m!} = 0,$$

otherwise the proof must be strictly geometric and must uncover the geometric meaning of every term in the series.

SOLUTION. Consider an arc x of a unit circle (Fig. 1); we divide it into n equal segments, and inscribe there a polygonal line $A_0A_1 \dots A_n$, which we call line D_1.

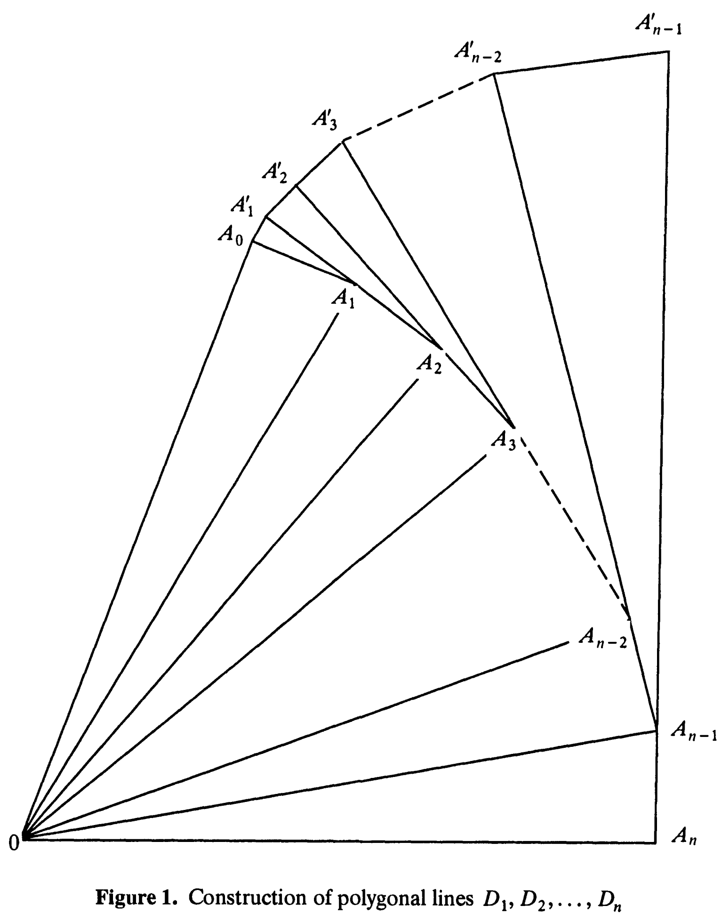


Figure 1. Construction of polygonal lines $D_1, D_2, \dots, D_n$

Then we lengthen each segment ($A_2A_1, A_3A_2 \dots A_nA_{n-1}$) in such a way that

$$A_1A_1' = A_1A_0 = 2\sin\frac{x}{2n}$$

$$A_2A_2' = A_0A_1 + A_1A_2 = 4\sin\frac{x}{2n}$$

$$\vdots$$

$$A_{n-1}A_{n-1}' = 2(n-1)\sin\frac{x}{2n}$$

Therefore, each segment is enlarged by adding the sum of the lengths of all the preceding segments (from the point A_0). Now, we connect points $A_0, A_1'A_2', \dots, A_{n-1}'$ and obtain a new polygonal line, which we call line D_2. We repeat the same procedure with the line D_2 and obtain line D_3, etc., up to the line D_n, which consists of one segment. Let us find the length of each line (designating them as L_i):

$$\begin{aligned}
L_1 &= 2\sin\frac{x}{2n}(1+1+\cdots+1) = 2n\sin\frac{x}{2n},\\
L_2 &= \left(2\sin\frac{x}{2n}\right)^2(1+2+\cdots+n-1) = 4\sin^2\frac{x}{2n}\frac{n(n-1)}{2},\\
&\vdots\\
L_i &= \left(2\sin\frac{x}{2n}\right)^i\left[\binom{i-1}{i-1}+\binom{i}{i-1}+\cdots+\binom{n+1-i}{i-1}\right] = \left(2\sin\frac{x}{2n}\right)^i\binom{n}{i}\\
&\vdots\\
L_n &= \left(2\sin\frac{x}{2n}\right)^n.
\end{aligned} \tag{1}$$

In order to prove the set of equations (1), it is sufficient to note that the triangles obtained during construction of polygonal lines are similar; we also have to use the recurrence

$$\binom{j}{i}+\binom{j}{i+1}=\binom{j+1}{i+1}.$$

Consider the limit

$$\begin{aligned}
L_1^0 &= \lim_{n\to\infty} L_1 = x,\\
L_2^0 &= \lim_{n\to\infty} L_2 = \frac{x^2}{2!},\\
L_3^0 &= \lim_{n\to\infty} L_3 = \frac{x^3}{3!},\\
&\vdots\\
L_m^0 &= \lim_{n\to\infty} L_m = \frac{x^m}{m!}.
\end{aligned} \tag{2}$$

The first polygonal line tends to an arc, the second one to its involute, the third one to the involute of the first involute, etc. We designate these arcs as $D_1^0, \dots, D_m^0 \dots$ respectively (Fig. 2). It is clear from the construction that $A_1A_2 =$

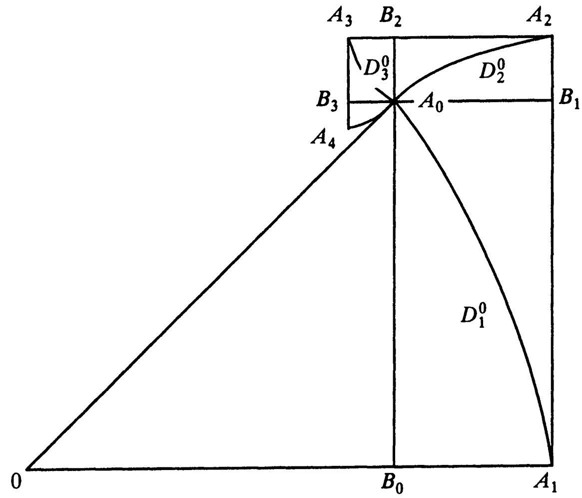


Figure 2. Arc D_1^0 and a sequence of involutes $D_2^0, D_3^0, \ldots$.

L_1^0, $A_2A_3 = L_2^0$, etc. But from Fig. 2 it can be directly seen that

$$\sin x = A_0B_0 < L_1^0 = A_1A_2 = x,$$

or, more exactly,

$$\sin x = A_1A_2 - B_1A_2 = x - B_1A_2 \equiv x - A_0B_2 > x - A_3A_4 = x - L_3^0 = x - \frac{x^3}{3!}.$$

Continuing, we obtain for any $m > 0$

$$x - \frac{x^3}{3!} + \cdots - \frac{x^{4m-1}}{(4m-1)!} < \sin x < x - \frac{x^3}{3!} + \cdots + \frac{x^{4m+1}}{(4m+1)!}. \tag{3}$$

When $m \to \infty$, the difference between the upper and lower bounds in (3) tends to zero and we obtain the expansion of $\sin x$ into series.

In the same manner we can obtain the expansion of $\cos x$ into series, since in Figure 2

$$\cos x = OB_0 = 1 - B_0A_1 = 1 - A_0B_1 > 1 - A_2A_3 = 1 - L_2^0 = 1 - \frac{x^2}{2!},$$

$$\begin{aligned}\cos x = OB_0 &= 1 - B_0A_1 = 1 - A_2B_2 = 1 - A_2A_3 + B_2A_3 \\ &= 1 - \frac{x^2}{2!} + B_2A_3 < 1 - \frac{x^2}{2!} + L_4^0 = 1 - \frac{x^2}{2!} + \frac{x^4}{4!},\end{aligned}$$

etc., that is, for any $m > 0$,

$$1 - \frac{x^2}{2!} + \cdots - \frac{x^{4m-2}}{(4m-2)!} < \cos x < 1 - \frac{x^2}{2!} + \cdots + \frac{x^{4m}}{(4m)!}. \tag{4}$$

1464 Stanford
Irvine, CA 92612
V1L2G@AOL.COM

UNSOLVED PROBLEMS

Edited by: **Richard Nowakowski & Richard Guy**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Nowakowski, Department of Mathematics & Statistics & Computing Science, Dalhousie University, Halifax NS, Canada B3H 3J5.

What is a Collineation of the Integer Plane?

J. W. (Michael) Lorimer

The idea for this paper was inspired by a closely related article submitted by Raymond Killgrove.

§1. AFFINE PLANES OVER SKEWFIELDS AND THEIR COLLINEATIONS. If D is a skew field (i.e., a non-commutative field) then the affine plane over D, $\mathscr{A}(D)$, has D^2 as points; the cosets $\alpha + \beta D$ where $\alpha \in D^2$ and βD is a free direct summand (i.e., $\beta \in D^2 \setminus \{0\}$) as lines; and parallelism defined by

$$\alpha + \beta D \| \delta + \tau D \Leftrightarrow \beta D = \tau D.$$

Of course, $\alpha + \beta D \| \delta + \tau D \Leftrightarrow \alpha + \beta D$ and $\delta + \tau D$ are equal or disjoint.

A collineation of $\mathscr{A}(D)$ is defined classically as a bijection ϕ: $D^2 \to D^2$ that maps collinear points to collinear points. A more modern approach first defines a *homomorphism* of $\mathscr{A}(D)$ (or any abstract affine plane, for that matter) as a pair $\phi = (\phi_1, \phi_2)$ where ϕ_1 maps points to points, ϕ_2 maps lines to lines and incidence and parallelism are preserved (i.e., $P \in \alpha + \beta D \Rightarrow \phi_1(P) \in \phi_2(\alpha + \beta D)$ and $\alpha + \beta D \| \delta + \tau D \Rightarrow \phi_2(\alpha + \beta D) \| \phi_2(\delta + \tau D)$). Then, a *collineation* of $\mathscr{A}(D)$ is a homomorphism $\phi = (\phi_1, \phi_2)$ so that ϕ_1 and ϕ_2 are bijections and $\phi^{-1} = (\phi_1^{-1}, \phi_2^{-1})$ is also a homomorphism. (See [5] and [2] for a comparison of these two approaches.)

It is a startling result of Corbas (see [5] and [2]) that a homomorphism $\alpha = (\alpha_1, \alpha_2)$ is already a collineation if α_1 is surjective! The fundamental theorem of affine geometry then says a collineation α is composed of a translation $\tau_A(X \rightsquigarrow X + A)$ and a nonsingular semi-linear transformation L: $D^2 \to D^2$.

§2. RUMINATIONS ON RING GEOMETRIES AND THEIR COLLINEATIONS. Some *natural questions leap out* (or perhaps they stroll): if we replace our skewfield D by an associative ring R with $0 \neq 1$ then what do we mean by the affine plane

over R, what is a collineation of this plane, and how can it be represented algebraically? In particular, we can ask this question about the *integer plane*, that is, the affine plane over $\mathbb{Z}$.

The immediate response would seem to be, well, just replace D by R. But wait! Things are not quite as they seem. For example, after we discover that for $\beta = (\beta_1, \beta_2) \in R^2$, βR is a free direct summand of R^2 if and only if there is a $\delta = (\delta_1, \delta_2) \in R^2$ so that $\beta_1 \delta_1 + \beta_2 \delta_2 = 1$ (i.e., β is *unimodular*) then even for the integers, the two lines $(1,1)\mathbb{Z}$ and $(0,1) + (-1,1)\mathbb{Z}$ are clearly not parallel but they are disjoint! Another tricky situation arises when we take a pair of points $(P, Q)(P \neq Q)$ and seek a line through P and Q. In this instance, we may have no lines through P and Q or infinitely many. In order to understand this non-classical behaviour of points and lines we require some way to ascertain when a pair (P, Q) lies on at most one or exactly one line. It is necessary to discuss these issues next, before we can determine how a collineation should act on points and lines. See **[7]** and **[10]** for details and other references.

§3. THE INTEGER PLANE AND OTHER AFFINE RING PLANES. As suggested by our preceding discussion, the *integer plane*, $A(\mathbb{Z})$, is the following structure:

- Points are all elements of $\mathbb{Z}^2$.
- Lines are all cosets $\alpha + \beta\mathbb{Z}$ with $\alpha, \beta \in \mathbb{Z}^2$ and $\beta = (\beta_1, \beta_2)$ so that β_1 and β_2 are relatively prime; in module language, β is *unimodular*.

Let $\mathscr{U}_m(\mathbb{Z})$ be the unimodular elements of $\mathbb{Z}^2$.

Two points P and Q are *non-neighbours* or *distant* (and we write $P \not\approx Q$) if $P - Q \in \mathscr{U}_m(\mathbb{Z})$ and are *neighbours* if $P \approx Q$. Then, two non-neighbours P and Q lie on the line $P + (P - Q)\mathbb{Z}$ and on no other line.

We can, of course, replace $\mathbb{Z}$ by any associate ring R with $0 \neq 1$. But, in this general setting we may choose a special non-empty subset $B_2 \subseteq R^2$ called a *Barbillian domain* of R^2; that is B_2 has the properties:

(B1) Every $\alpha \in B_2$ belongs to a basis of R^2 contained in B_2.
(B2) If $\{\alpha_1, \alpha_2\}$ is a basis of R^2 contained in B_2, then $\alpha_1 + \alpha_2 a \in B_2$ for all $a \in R$.

The set $B_2^{\max}(R) = \{\alpha \in R^2 | \alpha$ belongs to a basis of $R^2\}$ is obviously a Barbillian domain of R^2 and it contains all other Barbillian domains of R^2.

Now let B_2 be a Barbillian domain in R^2. The *affine Barbillian ring plane* over R and B_2, $\mathscr{A}(R, B_2)$, is defined as follows:

- Points are all elements of R^2.
- Lines are all subsets $\alpha + \beta R$ with $\alpha \in R^2$ and $\beta \in B_2$,
- Two lines are parallel $(\alpha + \beta R \| \gamma + \delta R)$ if $\beta R = \delta R$.

Then, as with $\mathbb{Z}$, two points P, Q are *non-neighbours* or *distant* (and we write $P \not\approx Q$) if $P - Q \in B_2$ and are *neighbours* if $P \approx Q$. Two non-neighbours P and Q lie on the line $P + (P - Q)R$ and this line is unique for any pair of non-neighbours if and only if R is a *two-sided units ring*, i.e., $ab = 1 \Rightarrow ba = 1$. Two lines l, m are neighbours $(l \approx m)$ if $L \in l$ implies that $L \approx M$ for some $M \in m$.

Now, how is $\mathscr{A}(R, B_2)$ related to the plane defined using the unimodular elements, $\mathscr{U}_m(R)$, of R as we did for the integer plane?

A finitely generated right R-module P is *stably* free of type m in rank n if $P \oplus R^m \simeq R^n$. A ring R is 2-hermite if it has invariant basis number 2 [$R^m \simeq R^2 \Rightarrow m = 2$] and stably free modules of type $m \leq 2$ in rank 2 are free [$r \leq m \leq 2$, $P \oplus R^r \simeq R^m \Rightarrow P \simeq R^{m-r}$]. Any commutative ring (and hence $\mathbb{Z}$) is 2-hermite.

If E_{ij} is the 2×2-matrix over R with 1 in the (i, j)-position and zeros elsewhere, then the group generated by the matrices $I_2 + aE_{ij}$ ($i \neq j$ and $a \in R$) and the invertible diagonal matrices, is $GE_2(R)$. R is a GE_2 ring if and only if $GE_2(R) = GL_2(R)$, where $GL_2(R)$ is the general linear group of R^2. Now any euclidean ring, and hence Z, is a GE_2 ring.

The preceding discussion now allows us to state the following important results:

($*$) For any Barbillian domain B_2 of R, $B_2 \subseteq \mathscr{U}_m(R^2)$. ([**7**, 12.4 (ii)])

($**$) If R is 2-hermite, then $B_2^{\max}(R) = U_m(R^2)$. ([**10**; 2.2])

($***$) R is a GE_2 ring if and only if $B_2^{\max}$ is the only Barbillian domain in R^2. ([**10**; 2.8])

As we have seen, $\mathbb{Z}$ is 2-hermite and a GE_2 ring, and so $\mathscr{U}_m(\mathbb{Z}^2)$ is the only Barbillian domain of $\mathbb{Z}^2$ and thus describes the unique Barbillian plane $\mathscr{A}(\mathbb{Z})$, over $\mathbb{Z}$.

§4. THE COLLINEATION GROUP OF THE INTEGER PLANE. Motivated by the notion of a homomorphism in ([**5**], [**2**]) and ([**4**], [**8**], and [**9**]) we call a pair $\phi = (\phi_1, \phi_2)$ a *homomorphism* of $\mathscr{A}(R, B_2)$ if ϕ_1 maps R^2 to R^2, ϕ_2 maps lines to lines, and the following properties hold:

(H1) ϕ preserves incidence, i.e., $P \in \alpha + \beta R \Rightarrow \phi_1(P) \in \phi_2(\alpha + \beta R)$.

(H2) ϕ preserves parallelism, i.e., $\alpha + \beta R \| \delta + \gamma R \Rightarrow \phi_2(\alpha + \beta R) \| \phi_2(\delta + \gamma R)$.

Now, what do we ask our homomorphism to do about $\approx$ and $\not\approx$? In our general presentation, $\not\approx$ appears via B_2 and so it seems a natural geometric operation for a homomorphism to preserve. However, for *local rings* (i.e., the non-units form an ideal J and then the unique Barbillian domain is $L^2 \setminus J^2$) that describe the *affine Klingenberg planes*, it is more natural to have a homomorphism preserve $\approx$. ([**6**] or [**1**])

We thus call a homomorphism $\phi = (\phi_1, \phi_2)$ *distant-preserving* or *neighbour preserving* if it preserves $\not\approx$ or $\approx$ respectively, i.e., $P \not\approx Q \Rightarrow \phi_1(P) \not\approx \phi_1(Q)$ and $l \not\approx m \Rightarrow \phi_2(l) \not\approx \phi_2(m)$. A *collineation* of $\mathscr{A}(R, B_2)$ is then a homomorphism $\phi = (\phi_1, \phi_2)$ that preserves $\not\approx$, and ϕ_1 and ϕ_2 are bijections so that $\phi^{-1} = (\phi_1^{-1}, \phi_2^{-1})$ is also a homomorphism that preserves $\not\approx$. Note that we can replace $\not\approx$ by $\approx$ and obtain an equivalent definition of a collineation. Of course, we can also define a collineation as a bijection ϕ_1: $R^2 \to R^2$ so that l is a line $\Leftrightarrow$ $\phi_1[l]$ is a line, $P \not\approx Q \Leftrightarrow \phi_1(P) \not\approx \phi_1(Q)$, and $l \| m \Leftrightarrow \phi_1[l] \| \phi_1[m]$.

Clearly, the set of collineations, C, of $\mathscr{A}(R, B_2)$ is a group. The subgroup $T(R^2)$, of translations $T\alpha_0$: $\alpha \rightsquigarrow \alpha + \alpha_0$, is the translation group; and $C_0 = \{\phi \in C | \phi(0) = 0\}$, the stabilizer of 0, is also a subgroup of C. Then, C is the semidirect product $C = C_0 \times T(R^2)$. Now, one can prove the following:

Fundamental Theorem. *C_0 is a subgroup of $\Gamma L_2(R)$ [the group of semilinear bijections of R^2], and $C_0 = \Gamma L_2(R)$ if and only if $B_2 = B_2^{\max}$.* ([**7**], [**10**])

Since in the integer plane $\mathscr{U}_m(R^2)$ is the only Barbillian domain and $\mathbb{Z}$ has no proper automorphisms, the collineations of the integer plane are

$C = GL_2(\mathbb{Z}) \cdot T(\mathbb{Z}^2)$, and $C_0 = \Gamma L_2(R)$, *where* $\mathrm{GL}_2(\mathbb{Z})$ *is the general linear group over* $\mathbb{Z}$ *and* $\mathrm{GL}_2(\mathbb{Z}) = \Gamma\mathrm{L}_2(\mathbb{Z}) = \mathrm{C}_0$.

§5. WHAT IS THE SIMPLEST DEFINITION FOR A COLLINEATION OF THE INTEGER PLANE? Now, for ordinary affine planes ([**5**] and [**2**]), and more generally for affine Klingenberg planes (and thus affine planes over local rings ([**3**] and [**1**])) we can simplify the definition of a collineation by observing the following results:

(I) A homomorphism preserves $\not\approx$ if and only if there are three points P, Q, R whose images P', Q', R' form a non-degenerate triangle, i.e., the vertices P', Q', R' are mutually non-neighbouring and each vertex is non-neighbouring to the opposite side.

In this case we call our homomorphism *non-degenerate* (or *full*). This just means that the image plane is also a Klingenberg plane. There are homomorphisms that preserve neither $\not\approx$ nor $\approx$ and there are non-degenerate homomorphisms that do not preserve $\approx$.

(II) If $\phi = (\phi_1, \phi_2)$ is a homomorphism and ϕ_1 is surjective then ϕ preserves both $\approx$ and $\not\approx$. In this case we call ϕ an *eumorphism.*

(III) A homomorphism $\phi = (\phi_1, \phi_2)$ is a collineation if and only if ϕ_1 is a bijection.

Now local rings have control over their inverses (i.e., if L is local then for each $x \in L$, either x or $1 - x$ is invertible) while $\mathbb{Z}$ has no invertible elements except ± 1. Thus, we can ask *is it possible to simplify the definition of a collineation for the integer plane*?

We address this query via the following natural problems about homomorphisms of $\mathscr{A}(\mathbb{Z})$.

Problem (1): What geometric conditions ensure that a homomorphism preserves $\not\approx$ or $\approx$? If ϕ preserves $\approx$ ($\not\approx$) does it preserve $\not\approx$ ($\approx$)?

In [**4**] a homomorphism $\phi = (\phi_1, \phi_2)$ between abstract Barbillian planes is called *full* if for all points X, Y there is a line l so that $\phi_1(X) \not\approx \phi_2(l)$ and $\phi_1(Y) \not\approx \phi_2(l)$. This notion of fullness for homomorphisms between abstract Klingenberg planes is easily seen to be equivalent to non-degeneracy. Algebraic answers to problem (1) were then given for projective Barbillian planes over rings of stable rank 2. Note: $\mathbb{Z}$ has stable rank 3.

Problem (2): If $\phi = (\phi_1, \phi_2)$ is a homomorphism and ϕ_1 is surjective, then does ϕ preserves $\not\approx$ and/or $\approx$?

Problem (3): Is a homomorphism $\phi = (\phi_1, \phi_2)$ a collineation if and only if ϕ_1 is bijective?

Problem (4): Is a bijection ϕ: $\mathbb{Z}^2 \to \mathbb{Z}^2$ that maps lines to lines a collineation?

Of course, we could also ask these questions for affine ring planes $\mathscr{A}(R, B_2)$, where R might be 2-hermite or stable rank 2. Even more generally, we could ask these questions about abstract Barbillian planes (see [**7**] and [**10**]).

REFERENCES

1. Phyrne Bacon, An introduction to Klingenberg planes, vol. I, II, III published by the author (1976–1979).
2. T. Bisztriczky and J. W. (Michael) Lorimer, Homomorphisms of affine spaces, *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* **65** (1995), 283–292.
3. T. Bisztriczky and J. W. (Michael) Lorimer, Homomorphisms and translations in affine Klingenberg spaces (in preparation).
4. J. C. Ferrar and F. D. Veldkamp, Neighbour preserving homomorphisms between projective ring planes, *Geom. Dedicata*, **18** (1985) 11–33.
5. J. W. (Michael) Lorimer and M. Buchholz, Homomorphisms of affine and projective planes, *Math. Mag.*, **55** (1982) 195–204.
6. J. W. (Michael) Lorimer, Affine Hjelmslev rings and planes, *Annals of Discrete Math.*, **37** (1988) 265–276.
7. F. D. Veldkamp, Geometry over rings, Chapter 19 of *Handbook of Incidence Geometry*, edited by F. Burkenhout, Elsevier Science B. V. (1995).
8. F. D. Veldkamp, Distant-preserving homomorphisms between projective ring planes, *Nederl. Akad. Wetensch. Proc. Ser. A88* (= *Indag. Math. 47*) (1985) 455–459.
9. F. D. Veldkamp, Incidence preserving mappings between projective ring planes, *Ned. Akad. Wetensch. Proc. ser. A88* (= *Indag. Math. 47*) (1985) 443–453.
10. F. D. Veldkamp, *n*-Barbillian domains, *Resultate Math.* **23** (1993) 177–200.

Department of Mathematics
University of Toronto
Toronto, Ontario M5S 1A1
Canada

Mathematics was very easy to me... My idea then was to get through the course, secure a detail for a few years as assistant professor of mathematics at the Academy, and afterwards obtain a permanent position as professor in some respectable college; but circumstances always did shape my course different from my plans.

Ulysses S. Grant, "Personal Memoirs of U. S. Grant", Vol. 1,
Charles L. Webster & Co., New York, 1885, pp. 39-40.

Contributed by Stephen B. Maurer, Swarthmore College

THE AUTHORS

ROBERT G. BARTLE was born in Kansas City, Missouri. He made his undergraduate studies at Swarthmore College with Arnold Dresden and H. W. Brinkmann, and his graduate studies at the University of Chicago with L. M. Graves (and many others). He worked at Yale University with Nelson Dunford, J. T. Schwartz, and W. G. Bade. From 1955–1990, he was on the staff of the University of Illinois, Urbana-Champaign, and served as Executive Editor of Mathematical Reviews (1976–1978, 1986–1990). Since 1990, he has been at Eastern Michigan University. His primary mathematical interests are in real analysis, integration theory, and functional analysis.

KATHY B. HAMRICK received her Ed.D. in Mathematics Education at the University of Georgia in 1976. She teaches mathematics at Augusta College in Augusta, Georgia. Her research interests have ranged from Elementary Teacher Education to Computer Science; she is the co-author of a now obsolete computer science text. She is a calculator nut and her current research interest is the history of the electronic hand-held calculator. She is a member of the International Association of Calculators and has a collection of over 100 old calculators.

RICHARD D. NEIDINGER is an Associate Professor of Mathematics at Davidson College in North Carolina. He moved there in 1984 after earning a B.A. from Trinity University in San Antonio and a M.A. & Ph.D., specializing in functional analysis, from the University of Texas at Austin. He enjoys using computers in numerical methods, fractals, and chaos to explore interesting topics that are accessible to undergraduates. In 1990, he received the George Pólya Award for "Automatic Differentiation and APL" in the *College Mathematics Journal.*

R. JOHN ANNEN, III is a system software analyst for the Forsyth County Government in Winston-Salem, NC. His article developed from his project as an undergraduate mathematics major at Davidson College. Since graduation in 1990, he has continued to pursue this and other mathematical interests. In addition, he chairs the Forsyth County Free-Net Project, Inc., a nonprofit organization committed to improving electronic access to information; he is a volunteer E.M.T./Rescue Technician with the Winston-Salem Rescue Squad; and he occasionally finds time to play electronic music.

STEPHEN BEDDING was born in Epsom, England, and educated at Durham and London Universities, where he majored in theoretical physics. He has held research posts at the Max Planck Institut in Munich and at English universities in Newcastle, Durham, London, and Southampton. During this time he published in the area of supergravity. He was subsequently a lecturer at Papua New Guinea University of Technology, where he became involved in a project to save an important leatherback turtle nesting beach. He is now senior lecturer in applied mathematics at La Trobe University in Melbourne, where he works on nonlinear dynamical systems when he is not playing the piano.

KEITH BRIGGS was born in Cambridge, England and educated at Adelaide University, Australia, where he majored in theoretical physics. He has held research and teaching posts in physics and applied mathematics at four universities in Adelaide, Wales, and Melbourne. During this time he published in the areas of chaos experiments, relativistic electrodynamics, Feigenbaum universality, discrete dynamical systems, colloid simulations, nonlinear time series, bioelectromagnetics, statistical mechanics, and random walk theory. Keith is now at the University of Western Australia in Perth, where he works on nonlinear dynamical systems when he is not playing the harpsichord.

JOHN STILLWELL was educated at Melbourne University and MIT, and has taught at Monash University since 1970. He has written several books, the most recent being *Elements of Algebra* (Springer-Verlag, 1994) and forthcoming translations of works by Beltrami, Klein, Poincare, and Dedekind. As these titles suggest, his interests are in algebra, geometry, and number theory, but occasionally he enjoys revisiting the logic and set theory he studied in graduate school.

KRISHNASWAMI ALLADI, from Madras, India, is Professor of Mathematics at the University of Florida since 1989. He was born in 1955. His undergraduate research in number theory attracted the attention of Paul Erdős. This led to a Chancellor's Fellowship at UCLA for a Ph.D. under Ernst Straus. After receiving the Alumni Medal for the best thesis in 1978, he was Hildebrandt Assistant Professor at Michigan, Associate Professor at MATSCIENCE in Madras, and Visiting Member at the Institute for Advanced Study, Princeton. He also held visiting positions at Texas, Hawaii, and Penn. State. His research is in analytic number theory, and more recently, in the theory of partitions and q-series. He has also won a teaching award at Florida. He is Editor-in-Chief of *The Ramanujan Journal*, a new international journal to be launched by Kluwer in 1997.

Storefront on 41st Road in Main Street, Flushing, NY... Contributed by Eizo Nishiura, Queensboro Community College.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Two copies of all items needed to evaluate the problem should be sent. A third copy of the the problem and solution is often useful; please include one if possible.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before March 31, 1997. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10543. *Proposed by Yuanan Diao, Kennesaw State College, Marietta, GA.*

Given an odd number of intervals, each of unit length, on the real line, let S be the set of numbers that are in an odd number of those intervals. Show that S is a finite union of disjoint intervals of total length at least 1.

10544. *Proposed by L. K. Jones, Shippensburg University, Shippensburg, PA.*

Prove that there is a finite set Q of odd primes with the following property. If $S = \{p_1, p_2, \ldots, p_k\}$ is a set of odd primes with $k \geq 2$ such that $Q \not\subseteq S$, then there exists an odd prime $q \notin S$ such that

$$\prod_{\lambda=1}^{k} (p_\lambda + 1) \equiv 0 \pmod{q+1}.$$

10545. *Proposed by Joaquín Gómez Rey, I. B. "Luis Buñuel", Alcorcón (Madrid), Spain.*

A set consisting of n men and n women is partitioned at random into n disjoint pairs of people. What are the expected value and variance of the number of male-female couples that result.

10546. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.*

Let m and n be integers, $n \geq m > 0$. Prove that, if q is an integer in the range $n - m < q \leq n$,

$$\sum_{k \equiv q \pmod 2} \binom{(k+q)/2 - 1}{k-1}\binom{n - (k+q)/2}{m-k} \equiv \binom{n}{m} \pmod 2.$$

10547. *Proposed by Dan Sachelarie, ICCE Bucharest, and Vlad Sachelarie (student), University of Bucharest, Bucharest, Romania.*

In the triangle ABC, let O be the circumcenter, H the orthocenter, and I the incenter. Prove that the triangle OHI is isosceles if and only if

$$\frac{a^3 + b^3 + c^3}{3abc} = \frac{R}{2r}.$$

10548. *Proposed by F. S. Cater, Portland State University, Portland, OR, and A. B. Thaheem, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.*

Let u and v be invertible elements of a noncommutative ring R such that

$$uvu^{-1} + u^{-1}vu = 2v, \tag{i}$$
$$u\left(vuv^{-1}\right) = \left(vuv^{-1}\right)u, \tag{ii}$$
$$v\left(uvu^{-1}\right) = \left(uvu^{-1}\right)v. \tag{iii}$$

Let R contain no nonzero nilpotent element; that is, if $x \in R$ and $x^2 = 0$, then $x = 0$.

(**a**) Prove that $vu = uv$.

(**b**)* Is the hypothesis on nilpotent elements necessary?

10549. *Proposed by Z. F. Starc, Vršac, Yugoslavia.*

Let $a_1, a_2, \ldots, a_n$ be positive real numbers and let $a = (a_1 + a_2 + \cdots + a_n)/n$. Prove that

$$\Gamma(a_1)^{\Gamma(a_1)} \cdot \Gamma(a_2)^{\Gamma(a_2)} \cdots \Gamma(a_n)^{\Gamma(a_n)} \geq e^{n(\Gamma(a)-1)}.$$

NOTES

(**10547**) We follow the usual convention that a, b, c are the lengths of the sides opposite vertices A, B, and C, respectively, and that R is the circumradius and r the inradius of ABC.

(**10548**) A paper giving some indication of the source of the problem is: A. B. Thaheem, "On certain decompositional properties of von Neumann algebras", *Glasgow Math. J.* 29 (1987), 177–179.

SOLUTIONS

A Lower Bound on the Identric Mean

10289 [1993, 185]. *Proposed by David M. Bloom, Brooklyn College of CUNY, Brooklyn, NY.*

For $x > 1$, consider the inequality

$$a\sqrt{x} + (1-a)\left(\frac{x+1}{2}\right) < e^{-1}x^{x/(x-1)}.$$

(**a**) If $a \geq 1/3$, show that the inequality holds for all $x > 1$.
(**b**) If $a < 1/3$, show that there is some $x > 1$ for which the inequality is false.

Solution by Heinz-Jürgen Seiffert, Berlin, Germany. The quantity $e^{-1}x^{x/(x-1)}$ is the identric mean of 1 and x, so that the result in (**a**) amounts to showing that $A(x,y) - I(x,y) < (1/3)\big(A(x,y) - G(x,y)\big)$, where $A(x,y)$, $G(x,y)$, and $I(x,y)$ are the Arithmetic, Geometric, and Identric means of two different positive numbers x and y. This result appears in J. Sándor, "A note on some inequalities for means", *Arch. Math.* 56 (1991), 471–473. Part (**b**) requires that this inequality be shown to be sharp. Chapter 8 of J. M. Borwein & P. B. Borwein, *Pi and the AGM*, Wiley, 1987 can serve as a guide to much of the literature on the subject. In particular, exercise 8.6.3 on page 264 of that book contains the essence of the desired result.

We prove (**b**) as a consequence of the easily verified identities

$$A(x,y) = G(x,y)\exp\left(\sum_{k=1}^{\infty}\frac{1}{2k}\left(\frac{x-y}{x+y}\right)^{2k}\right),$$

$$I(x,y) = G(x,y)\exp\left(\sum_{k=1}^{\infty}\frac{1}{2k+1}\left(\frac{x-y}{x+y}\right)^{2k}\right).$$

Let $u = (x-y)/(x+y)$. Then $A(x,y)/G(x,y) = 1+u^2/2+O(u^4)$ and $I(x,y)/G(x,y) = 1 + u^2/3 + O(u^4)$, so

$$I(x,y) - aG(x,y) - (1-a)A(x,y) = G(x,y)\big((3a-1)u^2/6 + O(u^4)\big).$$

Thus, the leading term is negative for $a < 1/3$.

The series in the formulas converge for $-1 < u < 1$, so these expressions are valid for all positive x and y.

We also note that the second of our formulas can be written

$$I(x,y) = G(x,y)\exp\left(\frac{A(x,y)}{L(x,y)} - 1\right),$$

where $L(x,y) = (x-y)/(\log x - \log y)$ when $x \neq y$.

Supplement: Another proof of (**a**) has appeared that relates these quantities to the Hölder means

$$H_p(x,y) = \left(\frac{x^p + y^p}{2}\right)^{1/p}$$

for $p \neq 0$. The "16th Austrian-Polish Mathematics Competition 1993" included the inequality

$$H_{2/3}(x,y) > \frac{G(x,y) + 2A(x,y)}{3}$$

for $x, y > 0$, $x \neq y$. Since $I(x,y) > H_{2/3}(x,y)$ for $x, y > 0$, $x \neq y$ is known from K. B. Stolarsky, "The power and generalized logarithmic means", this MONTHLY 87 (1980), 545–548, the desired result follows.

We have also obtained the following generalization of the contest problem.

Theorem. *(a) If $0 < p < 1/2$, then for all $x, y > 0, x \neq y$,*

$$G(x,y)^{1-p}A(x,y)^p < H_p(x,y) < (1-p)G(x,y) + pA(x,y).$$

(b) If $1/2 < p < 1$, then for all $x, y > 0, x \neq y$,

$$(1-p)G(x,y) + pA(x,y) < H_p(x,y).$$

Editorial comment. The statement of the problem encouraged solvers to solve for a, leading to a study of

$$R(x) = \frac{x + 1 - 2e^{-1}x^{x/(x-1)}}{x - 2\sqrt{x} + 1} = \frac{A(1,x) - I(1,x)}{A(1,x) - G(1,x)}$$

for $x \neq 1$. Note that the denominator is positive for $x \neq 1$, and the limit as $x \to 1$ has been evaluated in the solution to **(b)** given above. Most solvers obtained the result of **(a)** by showing $R(x)$ to be decreasing for $x > 1$. In all cases, the proof consisted of grouping terms in $R'(x)$ to allow known inequalities to be applied. We omit the details since these monotonicity results are stronger than the needed for this application. The result of Sándor cited above used a similar approach applied to $\left(2A(1,x) + G(1,x)\right) / I(1,x)$.

Solved also by R. Akhlagi & R. Dai, J. Anglesio (France), R. J. Chapman (U. K.), H. G. Killingbergtrø & I. Skau (Norway), K.-W. Lau (Hong Kong), J. Marengo, P. McCartney, A. D. Melas (Greece), A. Müller (Germany), B. E. Rhoades, P. J. Zwier, University of Wyoming Problem Circle, and the proposer. One incorrect solution was received.

Pseudo-aliquot Sequences

10323 [1993, 688]. *Proposed by David E. Penney and Carl Pomerance, University of Georgia, Athens GA.*

For a natural number n, let $t(n)$ be the sum of the divisors d of n in the range $1 \leq d < n$ with n/d being squarefree. Is there an integer n for which the sequence $n, t(n), t(t(n)), \cdots$ is unbounded?

Composite solution by Kevin Brown, University of South Alabama, Mobile, AL, and Charles Vanden Eynden, Illinois State University, Normal, IL. The answer is yes. We establish unboundedness when $n = 134850$.

In general, if $n = p_1^{a_1}p_2^{a_2}\cdots p_r^{a_r}$, where the $p_1, p_2, \ldots, p_r$ are distinct primes and the $a_1, a_2, \ldots, a_r$ are positive integers, then the divisors d of n with $1 \leq d < n$ and n/d squarefree are the numbers

$$\prod_{i=1}^{r} p_i^{a_i - \delta_i},$$

where each δ_i is 0 or 1 and not all δ_i are zero. Hence

$$t(n) = \prod_{i=1}^{r}\left(p_i^{a_i} + p_i^{a_i-1}\right) - n = n\left(\prod_{p|n}\left(1 + \frac{1}{p}\right) - 1\right), \qquad (*)$$

so that the ratio $t(n)/n$ depends only on the set of distinct primes dividing n. It follows that

$$t(m^k n) = m^k t(n) \text{ if } m|n \text{ and } k \text{ is a positive integer.} \qquad (**)$$

If $6|n$, then $(*)$ shows that $6|t(n)$ and that $t(n) > n$ unless n has the form $2^a 3^b$ for positive integers a and b. Since $t(2^a 3^b) = 2^a 3^b$, when $6|n$ the sequence $n, t(n), t^2(n), \cdots$ either goes monotonically to infinity or ends with a number of the form $2^a 3^b$ repeated indefinitely.

Letting $n_0 = 134850 = 2 \cdot 3 \cdot 5^2 \cdot 29 \cdot 31$ and $n_j = t(n_{j-1})$ for $j \geq 1$, $(*)$ yields

$$n_1 = 210750 = 2 \cdot 3 \cdot 5^3 \cdot 281, \qquad n_2 = 296850 = 2 \cdot 3 \cdot 5^2 \cdot 1979,$$
$$n_3 = 415950 = 2 \cdot 3 \cdot 5^2 \cdot 47 \cdot 59, \qquad n_4 = 620850 = 2 \cdot 3 \cdot 5^2 \cdot 4139,$$
$$n_5 = 869550 = 2 \cdot 3 \cdot 5^2 \cdot 11 \cdot 17 \cdot 31, \qquad n_6 = 1618770 = 2 \cdot 3 \cdot 5 \cdot 53959,$$
$$n_7 = 2266350 = 2 \cdot 3 \cdot 5^2 \cdot 29 \cdot 521, \qquad n_8 = 3371250 = 2 \cdot 3 \cdot 5^4 \cdot 29 \cdot 31 = 25n_0.$$

Since $n_0, n_1, \ldots, n_8$ are all divisible by 5, we readily deduce (using $(**)$ and induction) that $n_{j+8} = 25n_j$ for every nonnegative integer j. Hence, $n_{8k+r} = 5^{2k} n_r$ if $0 \leq r < 8$ and k is a nonnegative integer.

Editorial comment. All other solvers showed similarly that the sequence $\{t^i(9870)\}_{i=0}^{\infty}$ is unbounded and monotonic. Brief calculations show that $t^{19}(9870) = 27 \cdot 9870$. As in the above solution, this yields $t^{19k+r}(9870) = 3^{3k} t^r(9870)$ for $0 \leq r < 19$ and k a nonnegative integer. Only these two sequences were received as solutions. The proposers ask whether it is possible for the sequence $n, t(n), t^2(n), \cdots$ to be unbounded but not monotonic.

The popularity of the latter example was most likely due to the fact that the smallest positive integer n for which the sequence $n, t(n), t^2(n), \cdots$ is unbounded is $n = 318$ and that $t^{12}(318) = 9870$.

The aliquot sequence problem asks the same question as the present problem, except that the condition "n/d being squarefree" is dropped. The aliquot sequence problem is unsolved; for example, the status of the aliquot sequence beginning with 276 is unresolved. For further discussion see §B6 of Richard K. Guy, *Unsolved Problems in Number Theory*, second edition, Springer, 1994.

Solved also by D. Alvis, R. J. Chapman (U. K.), R. Holzsager, N. Jensen (Germany), O. P. Lossers (The Netherlands), R. Martin (student), A. N. 't Woord (The Netherlands), and the proposers.

A Convolution Square Root

10329 [1993, 689]. *Proposed by Gérard Letac, Université Paul Sabatier, Toulouse, France.*

Let $f(x)$ is a positive continuous function defined for $0 < x < 1$ such that, for all u with $0 < u < 1$, one has $\int_u^1 f(x)x(u/x)\,dx = u^{1/2}$. Prove that

$$f(x) = \sqrt{\frac{2x}{\pi(1-x^2)}}.$$

Solution I by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Let $x = e^{-y}$, $u = e^{-z}$, and $G(y) = f\left(e^{-y}\right)$. Also let $R(y)$ denote $(\pi \sinh y)^{-1/2}$. This leads to the equivalent problem of showing that

$$\int_0^z G(z-y)G(y)e^{-y}\,dy = e^{-z/2} \tag{1}$$

for $z > 0$ implies that $G(y) = R(y)$. We now solve this transformed problem.

If we denote the Laplace transform of G by g, then (1) is equivalent to

$$g(s)g(s+1) = \frac{1}{s+\frac{1}{2}}, \tag{2}$$

while elementary integration yields that the Laplace transform of R is

$$r(s) = \frac{1}{\sqrt{2}}\Gamma\left(\frac{s}{2}+\frac{1}{4}\right) \Big/ \Gamma\left(\frac{s}{2}+\frac{3}{4}\right),$$

which implies that

$$r(s)r(s+1) = \frac{1}{s+\frac{1}{2}}. \tag{3}$$

Since $G(y)$ and $R(y)$ are positive, both $g(s)$ and $r(s)$ are decreasing functions. Thus (2) implies

$$\frac{1}{\sqrt{s+\frac{1}{2}}} \le g(s) \le \frac{1}{\sqrt{s+\frac{3}{2}}},$$

and (3) implies similar inequalities on $r(s)$. Thus $g(s) \sim s^{-1/2}$ and $r(s) \sim s^{-1/2}$ as $s \to \infty$.

From (2) and (3) it then follows that

$$\frac{g(s)}{r(s)} = \frac{r(s+1)}{g(s+1)} = \frac{g(s+2)}{r(s+2)} = \cdots = \frac{g(s+2n)}{r(s+2n)} \to 1 \quad (n \to \infty).$$

Hence, $g(s) = r(s)$, which gives $G(y) = R(y)$, as required.

Solution II by Godfrey L. Isaacs, Hollywood, FL. The stated problem is the case $k = 1/2$ of the apparently more general problem of showing that the only solution, for arbitrary real k, of $\int_u^1 f(x)f(u/x)\,dx = u^k$ is $f(x) = (2/\pi)^{1/2}x^k \,/\, \left(1-x^2\right)^{1/2}$. However, the substitution $g(x) = x^{-k}f(x)$ shows that this follows from the case with $k = 0$. Thus, we show that

$$\int_u^1 g(x)g(u/x)\,dx = 1 \tag{$*$}$$

has only the solution $g(x) = (2/\pi)^{1/2} \,/\, \left(1-x^2\right)^{1/2}$

To show this, write

$$I = \int_w^1 g(w/u) \int_u^1 g(x)g(u/x)\,dx\,du$$

where $0 < w < 1$. Then $(*)$ gives that

$$I = \int_w^1 g(w/u)\,du = \int_w^1 g(t)\left(w/t^2\right)dt$$

with $t = w/u$. Reversing the order of integration in I (which is justified since the integrand is positive) gives

$$I = \int_w^1 g(x) \int_w^x g(w/u)g(u/x)\,du\,dx = \int_w^1 xg(x)\,dx$$

after substituting $u = xy$ in the inner integral and using $(*)$. Comparing these two values of I, it follows that

$$\int_w^1 xg(x)\,dx = w \int_w^1 g(x)x^{-2}\,dx.$$

The continuity of g allows this to be differentiated to obtain

$$g(w)\frac{1-w^2}{w} = \int_w^1 g(x)x^{-2}\,dx.$$

This shows that g is differentiable and $g'(w) \ / \ g(w) = w \ / \ (1 - w^2)$. Hence $g(w) = C(1 - w^2)^{-1/2}$, and $(*)$ gives that $C = \left(2/\pi\right)^{1/2}$.

Editorial comment. S. K. Rangarajan used the Mellin transform and considered other functions that could be used in place of $u^{1/2}$ while allowing solutions of a fairly simple type. Some examples are $u^{\alpha+1/2}$, $u^{\alpha+1/2}\left(1 - u^h\right)\nu - 1$ (with $\nu h = 2$), or $u^m P_n(u)$ with P_n the Legendre polynomial of degree n. The unique solution was obtained in these cases.

The Western Maryland College Problems group began as in Solution II and interchanged the order of integration in

$$M_k = \int_0^1 u^k \int_u^1 g(x)g(u/x)\,dx\,du$$

for nonnegative integers k. This allows a suitable constant multiple of $g(x)$ to be interpreted as a probability density function and leads to equations determining its moments.

Solved also by K. F. Andersen & W.-S. Young (Canada), A. Chakrabarti (India), R. J. Chapman (U. K.), D. A. Darling, M. Golomb, A. A. Jagers (The Netherlands), K.-W. Lau (Hong Kong), T. L. McCoy, W. A. Newcomb, S. K. Rangarajan (India), Western Maryland College Problems group, and the proposer. Six incorrect or incomplete solutions were received.

A Consequence of Everything

10330 [1993, 796]. *Proposed by R. Bruce Richter, Carleton University, Ottawa, Ontario, Canada, and Josef Širáň, Technical University of Bratislava, Bratislava, Slovakia.*

Let n and k be given positive integers. Define q, r, s, t to be the unique integers such that $n = qk + r = s(k+1) + t$, with $0 \le r < k$ and $0 \le t \le k$. Show that

$$\binom{q}{2}k + rq \ge \binom{s}{2}(k+1) + ts.$$

Solution I by GCHQ Problem Solving Group, Cheltenham, U.K. For fixed n, let k be a positive real variable. Define $q = q(k)$ and $r = r(k)$ by $n = qk + r$, where $q = \lfloor n/k \rfloor$ and $r \in [0, k)$. Define f by $f(k) = \binom{q}{2}k + rq$. It suffices to show that f is a nonincreasing function of k. When n/k is not an integer, $q(k)$ is constant in a neighborhood of k, and f is differentiable with $f'(k) = \binom{q}{2} - q^2$, which is negative for $k < n$ and vanishes above n. When n/k is an integer a, the function f is continuous, since the limits from the left and right are $\binom{a}{2}n/a$ and $\binom{a-1}{2}n/a + (n/a)(a-1)$, which are equal. The result follows.

Composite solution II by Bill Doran, California Institute of Technology, Pasadena, CA, and the editors. Define $f(k)$ as above. Arrange n squares into an array (Ferrer's diagram) with s rows of $k+1$ squares and a final row of t squares. The number of ways of picking two squares in the same column is $f(k+1)$, where the two terms arise from whether or not the pair uses the row of length t. We now remove the s squares in the last column (losing $\binom{s}{2}$ pairs) and place them successively on the bottom, gaining at least s pairs from each, until we complete q rows of k squares followed by a row with r squares. Now there are $f(k)$ pairs, and we have $f(k) - f(k+1) \ge s^2 - \binom{s}{2} = \frac{s^2+s}{2}$. Thus, $f(k) \ge f(k+1)$, with equality only if $s = 0$. This happens when $k \ge n$, in which case $f(k) = 0$.

Solution III by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Define $f(k)$ as above. For each positive integer k, $f(k) = k\left(\sum_{i=0}^{q-1} i\right) + rq = \sum_{j=1}^{n} \left\lfloor \frac{j-1}{k} \right\rfloor$. Hence $f(k)$ is the number of integer lattice points (j, y) such that $1 \le j \le n$ and

$0 < y < j/k$. The set of points counted by $f(k)$ contains that counted by $f(k+1)$, so $f(k) \geq f(k+1)$.

Solution IV by the editors. The value $f(k)$ defined above is the number of edges in the complement of $T_{n,k}$, the complete k-partite graph with n vertices. Turán's Theorem states that $T_{n,k+1}$ is the unique maximum-sized n-vertex graph having no complete subgraph with $k+2$ vertices. Since this class includes $T_{n,k}$, we obtain $f(k) > f(k+1)$ for $k < n$.

Editorial comment. Most solvers used algebraic manipulations to obtain the inequality. For example, Richard Holzsager computed the exact difference $f(k) - f(k+1) = \big(q - s\big)\big(n - (q+s+1)k/2\big) + \binom{s+1}{2}$, which is nonnegative because $q \geq s$.

Solved also by J. Anglesio (France), R. Barbara (Lebanon), K. L. Bernstein, J. C. Binz (Switzerland), S. Byrd, D. Callan, J. Christopher, B. Correll, Jr. (student), P. L. Douillet (France), R. Geretschläger (Austria), R. Holzsager, R. D. Hurwitz, N. Komanda, H. Maharaj (student, South Africa), A. Pedersen (Denmark), R. M. Robinson, M. Vowe (Switzerland), K. S. Williams (Canada), A. N. 't Woord (The Netherlands), M. Zamaklar (student, Yugoslavia), National Security Agency Problems Group, and the proposers.

Multiply Perfect Numbers

10331 [1993, 796]. *Proposed by Carl Pomerance, University of Georgia, Athens, GA.*

Find all positive integers n such that $n!$ is multiply perfect; i.e., a divisor of the sum of its positive divisors.

Solution by Ulrich Everling, Kath. Universität, Eichstätt, Germany. The only solutions are $n = 1, 3, 5$. Moreover, for $n > 7$ the prime factor 2 has lower exponent in the sum of divisors of $n!$ than in $n!$ itself. For $n \leq 7$, the claim is checked explicitly using formula (1) below with $m = n!$. For $n > 7$, the key ingredient in the proof is Chebyshev's prime number theorem, as refined by Rosser and Schoenfeld.

Throughout, p denotes a prime number and q denotes an odd prime number. If x is a nonzero rational number, we let $\nu_p(x)$ be the exponent of p in the prime factorization of the expression of x as a fraction in lowest terms (if p divides the denominator, then $\nu_p(x)$ is negative). As usual, let $\pi(x)$ be the number of primes less than or equal to x, for $x \in \mathbb{R}$.

For each positive integer m, the sum of positive divisors of m is

$$s_m = \prod_p \sum_{i=0}^{\nu_p(m)} p^i = \prod_p \frac{p^{\nu_p(m)+1} - 1}{p - 1}. \tag{1}$$

Note that the factor for $p = 2$ in s_m is always odd. Furthermore, $\nu_p(n!) = \sum_{k=1}^{\infty} \lfloor n/p^k \rfloor$. For each q and n, there is a unique odd number $u_q(n)$ and a unique nonnegative integer $r_q(n)$ such that $\nu_q(n!) + 1 = 2^{r_q(n)} u_q(n)$. It follows that

$$q^{\nu_p(n!)+1} - 1 = \left(q^{u_q(n)} - 1\right) \prod_{k=0}^{r_q(n)-1} \left(q^{2^k u_q(n)} + 1\right).$$

If u is odd, then $x^u \pm 1$ is an odd multiple of $x \pm 1$. Hence for all $x \in \mathbb{Z}$, we have $\nu_2(x^u - 1) = \nu_2(x-1)$ and $\nu_2(x^u + 1) = \nu_2(x+1)$. For $k > 0$, we have $\nu_2\big(q^{2^k} + 1\big) = 1$ because $q^2 \equiv 1 \pmod 4$. Therefore, if Q_n is the set of odd primes for which $r_q(n) > 0$, we have

$$\nu_2(s_{n!}) = \sum_{q \in Q_n} \big(r_q(n) - 1 + \nu_2(q+1)\big). \tag{2}$$

We bound the contributions to (2). From $2^{r_q(n)}u_q(n) - 1 = \nu_q(n!) < n\sum_{k=1}^{\infty} q^{-k} < n/(q-1)$, we obtain $n > (2^{r_q(n)}u_q(n) - 1)(q-1)$. For $0 < k < r_q(n)$, this implies $n > (2^{k+1}-1)(q-1) \geq 2^k q$. Therefore,

$$\sum_{q\in Q_n} (r_q(n) - 1) = \sum_{k>0} \#\{q : r_q(n) > k\} \leq \sum_{k>0} \pi(2^{-k}n).$$

Next consider $\nu_2(q+1)$. The factors $q+1$ for $q \in Q_n$ all occur in $(n+1)!$, and the quotient $\frac{(n+1)!}{\prod_{q\in Q_n}(q+1)}$ contains the factors $j+1$ such that $1 \leq j \leq n$ and j is not prime. Each of the $\lceil n/2\rceil - \pi(n) + 1$ odd values of j contributes $\nu_2(j+1) \geq 1$. Hence $\lceil n/2\rceil - \pi(n) + 1 \leq \nu_2\left(\frac{(n+1)!}{\prod(q+1)}\right) = \nu_2\big((n+1)!\big) - \sum \nu_2(q+1)$. With $\nu_2\big((n+1)!\big) = \nu_2(n!) + \nu_2(n+1)$, we obtain $\sum_{q\in Q_n} \nu_2(q+1) \leq \nu_2(n!) + \nu_2(n+1) - \lceil n/2\rceil + \pi(n) - 1$.

Applying these computations to (2) yields

$$\nu_2\left(\frac{s_n!}{n!}\right) < \sum_{k>0} \pi(2^{-k}n) + \nu_2(n+1) - \left\lceil \frac{n}{2} \right\rceil + \pi(n) \leq -\frac{n}{2} + \log_2(n+1) + \sum_{k\geq 0} \pi(2^{-k}n). \quad (3)$$

The terms with $2^k > n/2$ vanish, because $\pi(x) = 0$ for $x < 2$.

By the Prime Number Theorem, we can choose $c \in \mathbb{R}$ such that $\pi(x) < cx/\ln x$ for all $x > 1$. For each positive integer i such that $c < 2^i \ln 2 < n \ln 2$, we have

$$\begin{aligned}\sum_{k=i+1}^{\lfloor \log_2(n/2)\rfloor} \pi(2^{-k}n) &\leq (\log_2(n/2) - i)\pi(2^{-1-i}n) \\ &\leq (\log_2(n/2) - i)\frac{c\cdot 2^{-1-i}n}{\ln n - (i+1)\ln 2} = \frac{2^{-i}cn}{2\ln 2}.\end{aligned}$$

Substituting this into (3) and dividing by n yields

$$\frac{1}{n}\nu_2\left(\frac{s_n!}{n!}\right) \leq -\frac{1}{2} + \frac{\log_2(n+1)}{n} + \frac{2^{-i}c}{2\ln 2} + c\sum_{k=0}^{i} \frac{2^{-k}}{\ln n - k\ln 2}.$$

This bound is negative for sufficiently large n. In particular, it is negative when $c = 1.25506$, $i = 6$, and $n \geq 488$. This is checked numerically for $n = 488$, and the bound is a decreasing function of n. The remaining cases $7 < n < 488$ are easily computed using (2). The constant c appears in formula (3.6) on page 69 of J. B. Rosser and L. Schoenfeld, "Approximate formulas for some functions of prime numbers", *Illinois J. Math.* 6(1962), 64–94.

Editorial comment. The proposer gave a shorter proof by using a theorem of Schinzel that $2^m - 1$ has a prime factor at least $2m+1$ for $m \geq 13$, from A. Schinzel, "On primitive prime factors of $a^n - b^n$", *Proc. Cambridge Phil. Soc.* 58(1962), 555-562.

Solved also by the National Security Agency Problems Group and the proposer. Two incorrect solutions were received.

Another combinatorial identity

10332 [1993, 796]. *Proposed by Kiran Kedlaya, student, Harvard University, Cambridge, MA.*

If n and k are integers with $0 \leq k \leq n$, prove that

$$\binom{2n}{n+k} = \sum_j 2^{n-k-2j}\binom{n}{j}\binom{n-j}{j+k}.$$

Solution I by Tim Hesterberg, Franklin and Marshall College, Lancaster, PA. We prove that both sides count the same set. The number of ways to select $n-k$ people from n couples is $\binom{2n}{n-k} = \binom{2n}{n+k}$. For each such selection, there is some number j of complete couples selected. There are $\binom{n}{j}$ ways to choose these couples. The rest of the selected group consists of $n-k-2j$ people without their spouses. There are $\binom{n-j}{n-k-2j} = \binom{n-j}{j+k}$ ways to choose the couples from which these others are selected and 2^{n-k-2j} ways to choose one member from each such couple. Summing over j completes the count.

Solution II by J. C. Binz, University of Bern, Bern, Switzerland. The coefficient of x^{n-k} in the expansion of $(1+x)^{2n}$ is $\binom{2n}{n+k} = \binom{2n}{n-k}$. But

$$(1+x)^{2n} = \left(x^2 + (1+2x)\right)^n = \sum_j \binom{n}{j} x^{2j}(1+2x)^{n-j}$$
$$= \sum_j \binom{n}{j} x^{2j} \sum_i \binom{n-j}{i} 2^i x^i,$$

and the coefficient of x^{n-k} is

$$\sum_j \binom{n}{j}\binom{n-j}{n-k-2j} 2^{n-k-2j} = \sum_j \binom{n}{j}\binom{n-j}{j+k} 2^{n-k-2j},$$

which proves the identity.

Editorial remark: Several solvers noted the similarity between this problem and E3258 (this MONTHLY, [1988, 259; 1989, 651]). The two solutions parallel the two solutions printed here. The editorial comment following those solutions mentions the identity in this problem, describing it as "a special case of Vandermonde's identity". Other solvers noted that the present identity follows immediately from an identity on page 79 of Riordan, *Combinatorial Identities*. Michael Hauss considered the case in which n and k are allowed to be complex numbers, which he denoted by α and β, respectively, with $\Re\alpha > -1/2$ to guarantee convergence. The sum should then be taken over nonnegative integers j. In this generality, the identity is essentially a hypergeometric identity of Gauss.

Solved also by J. Anglesio (France), K. L. Bernstein, M. Bhargava (student), M. Brahm, A. E. Caicedo Núñez (student, Colombia), D. Callan, T. E. Chambers (student), R. J. Chapman (U. K.), K.-J. Chen (Taiwan), T. H. Crocker, B. Doran, J. W. Grossman, H. S. Gunaratne (Brunei), M. Hauss (Germany), A. A. Jagers (The Netherlands), L. Krussel & S. G. Penrice, J. H. Lee (student, Korea), J. H. van Lint (The Netherlands), O. P. Lossers (The Netherlands), H. Maharaj (student, South Africa), R. Martin (student), S. Pedersen (Denmark), K. Pimsamarn, H. Prodinger (Austria), D. Radcliff, E. Schmeichel, H.-J. Seiffert (Germany), A. Tissier (France), M. Vowe (Switzerland), K. S. Williams (Canada), A. N. 't Woord (The Netherlands), B. Zimmermann (Austria), Anchorage Math Solutions Group, GCHQ Problem Solving Group (U. K.), Western Maryland College Problems group, and the proposer.

A Disguised Zeta Function

10333 [1993, 797]. *Proposed by Michael Golomb, Purdue University, West Lafayette, IN.*

For a positive integer n with $2^k \le n < 2^{k+1}$, let $L(n) = 2^k$ ($k = 0, 1, 2, \ldots$). Let $S(n)$ be the sum of the binary digits of n.

(a) Evaluate $\sum_{n\ge 1} \frac{1}{L^2(n)S(n)}$.

(b) Show that $\sum_{n\ge 1} \frac{1}{L(n)S(n)}$ diverges.

(c) Show that $\sum_{n\ge 1} \frac{1}{L(n)S^{1+\delta}(n)}$ converges for every $\delta > 0$.

Solution by A. N. 't Woord, University of Technology, Eindhoven, the Netherlands. First observe that there are exactly $\binom{k}{l}$ integers n with $2^k \le n < 2^{k+1}$ and $S(n) = l + 1$. Also note that all terms are positive, so that we can change the order of summation. In parts (a) and (c) we will need the following identity for $|x| < 1$:

$$\sum_{k=l}^{\infty} \binom{k}{l} x^k = \frac{x^l}{l!} \frac{d^l}{dx^l} \left(\sum_{k=0}^{\infty} x^k \right) = \frac{x^l}{l!} \frac{d^l}{dx^l} \left(\frac{1}{1-x} \right) = \frac{x^l}{(1-x)^{l+1}}.$$

Solution of (a)

$$\sum_{n \ge 1} \frac{1}{L^2(n) S(n)} = \sum_{k=0}^{\infty} 2^{-2k} \sum_{l=0}^{k} \frac{1}{l+1} \binom{k}{l} = \sum_{l=0}^{\infty} \frac{1}{l+1} \sum_{k=0}^{\infty} \binom{k}{l} \left(\frac{1}{4} \right)^k$$

(where the terms of the inner sum are zero for $k < l$)

$$= \sum_{l=0}^{\infty} \frac{1}{l+1} \left(\frac{1}{4} \right)^l \left(\frac{4}{3} \right)^{l+1} = 4 \sum_{l=0}^{\infty} \frac{(1/3)^{l+1}}{l+1}$$
$$= -4 \log(1 - 1/3) = 4 \log(3/2).$$

Solution of (b)

$$\sum_{n \ge 1} \frac{1}{L(n) S(n)} = \sum_{k=0}^{\infty} 2^{-k} \sum_{l=0}^{k} \frac{1}{l+1} \binom{k}{l}$$
$$\ge \sum_{k=0}^{\infty} 2^{-k} \sum_{l=0}^{k} \frac{1}{k+1} \binom{k}{l} = \sum_{k=0}^{\infty} \frac{1}{k+1} = \infty.$$

Solution of (c)

$$\sum_{n \ge 1} \frac{1}{L(n) S^{1+\delta}(n)} = \sum_{k=0}^{\infty} 2^{-k} \sum_{l=0}^{k} \frac{1}{(l+1)^{1+\delta}} \binom{k}{l}$$
$$= \sum_{l=0}^{\infty} \frac{1}{(l+1)^{1+\delta}} \sum_{k=l}^{\infty} 2^{-k} \binom{k}{l} = \sum_{l=0}^{\infty} \frac{2}{(l+1)^{1+\delta}}$$

This convergence of this last expression is a familiar exercise. Its value is $2\zeta(1+\delta) < \infty$.

Editorial comment. Several readers considered $\sum_{n \ge 1} \frac{1}{L^{\alpha}(n) S^{\beta}(n)}$ for other values of α and β. The most complete solution was given by Alain Tissier who showed that for $\alpha, \beta \in \mathbb{R}$, the series converges if and only if ($\alpha > 1$) or ($\alpha = 1$ and $\beta > 1$).

Solved also by K. L. Bernstein, A. E. Caicedo Núñez (student, Colombia), D. Callan, R. J. Chapman (U. K.), C. Cooper, T. H. Crocker, P. L. Douillet (France), J. S. Frame, C. Georghiou (Greece), E. Hertz, R. Holzsager, O. P. Lossers (The Netherlands), R. Martin (student), C. A. Minh (student), A. Pedersen (Denmark), K. Schilling, N. C. Singer, A. Tissier (France), C. Vanden Eynden, M. Vowe (Switzerland), Anchorage Math Solutions Group, GCHQ Problem Solving Group (U. K.), National Security Agency Problems Group, The Citadel Problem Solving Group, Western Maryland College Problems group, and the proposer. Two solutions were received that gave an incorrect analysis of part (c).

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

In Search of Infinity, N. Ya. Vilenkin, Birkhäuser, 1995

Reviewed by **John Stillwell**

Not long ago, mathematics had "the crisis in intuition." In fact, this was exactly the title of an influential article by the Austrian logician Hans Hahn (1879–1934). Written around 1930, it reached the peak of its popularity around 1960, when an English translation of it was published in James R. Newman's anthology *The World of Mathematics*. The message of the article was that intuition cannot be trusted. Hahn discussed the shocks delivered to intuitive notions of space and time by discoveries such as noneuclidean geometry and the theory of relativity, but his main targets were the intuitive concepts of analysis: continuity, differentiability, area, and the concept of curve. With a well-chosen series of examples, he exploded the "obvious" ideas that a continuous curve must have tangents, and must have zero area.

His examples of curves without tangents, and that fill space, will be pretty familiar to today's readers. The same examples have been described and illustrated in a dozen recent books on chaos and fractals. The difference is that Hahn *denies* we can understand these examples through pictures, and insists they can be understood only through logic. Well, since he was a logician, he would say that, but distrust of pictures is still with us. I am often surprised by my more critical students, who nod when I explain a theorem geometrically, then say "Yes, but how do you prove it?" We have the technology to produce vivid, intricate, and mathematically accurate pictures, but it fails to carry conviction. Have we been brainwashed by Hahn's generation, or is there really something lacking in visual reasoning?

In the case of elementary geometry, we know what is lacking—Hilbert filled the gaps in his *Foundations of Geometry* at the turn of the century, and was not afraid to use pictures in the process. In effect, he found exactly how much logical support is needed for traditional visual proofs in geometry, so intuitive proofs can routinely be brought up to scratch. The reason the "crisis in intuition" is still with us, I suspect, is that the same has not been done for the infinite geometric constructions of analysis. We lack a language that accurately describes the infinite processes behind the various nondifferentiable and space-filling curves and surfaces. We need the language of set theory, whose whole purpose is the mathematical

description of infinity. If geometry is combined with set theory, intuition may perhaps be refined to the point where the "counterintuitive" results become intuitive again.

The great problem, for many geometrically-minded people, is that set theory is bizarre, paradoxical, and unrealistic. It is the last field they expect to be "intuitive," and the last field they wish to combine with geometry. This is where Vilenkin comes to the rescue. He makes an irresistible case for set theory in geometry, not only by using it to fill the gaps where intuition allegedly deserts us, but also by showing that set theory is as much *fun* as geometry.

To put set theory in its proper perspective, the book begins with an informal introduction to the history and philosophy of infinity. We learn of the tension between continuous and discrete in ancient Greek mathematics, Zeno's paradoxes, Aristotle's five reasons for believing in infinity (the most important, in his opinion, was that "there are no bounds to thought") and the controversy over actual versus potential infinity. Astronomy also made some provocative contributions to the debate, with Bruno's claim that the universe is infinite, Olbers' paradox (if there are infinitely many stars, why isn't the night sky filled with light?), Hubble's discovery of expansion, and the infinite dilation of time experienced by a person falling into a black hole. This leads Vilenkin into a discussion of curved space and hyperbolic geometry, which is well done though possibly a little off the subject.

At any rate, by page 30 the stage is set. Vilenkin has convinced us that infinity is a burning question; we are eager to learn more about it, and it is time to tackle the subject mathematically. Fortunately, the mathematics of infinity makes every bit as good a story as its history and philosophy, if properly handled, and Vilenkin is a master of mathematical story-telling.

I expect that many readers of the *Monthly* will have heard of "Hilbert's Hotel," the hotel with infinitely rooms that can accept more guests even when it is full. Vilenkin renames it the Hotel Cosmos, and tops all previous Hilbert Hotel stories with a rollicking series of science fiction adventures. First the full hotel has to accommodate one extra guest, then a party of infinitely many, then infinitely many infinite parties. In completely engrossing style, Vilenkin uses these three scenarios to explore the basic arithmetic of countable sets. He exploits their narrative and comic possibilities to the hilt, seizing the reader's attention until the mathematics is crystal clear, at which point the three scenarios condense to the three equations $|\mathbb{N}| + 1 = |\mathbb{N}|$, $|\mathbb{N}| + |\mathbb{N}| = |\mathbb{N}|$ and $|\mathbb{N}| \times |\mathbb{N}| = |\mathbb{N}|$. Having seen vividly what these equations actually mean, there is no shock to the intuition—the equations express properties of the countable set $\mathbb{N}$ that by this stage are obvious.

Vilenkin has tremendous fun with the $|\mathbb{N}| \times |\mathbb{N}| = |\mathbb{N}|$ scenario, using it to discuss various ways of assigning guests to rooms, and judging the best to be the Cantor pairing function. (This function sends the nth guest in the mth party to room $(n-1)^2 + m$ if $n \geq m$ and to room $m^2 - n + 1$ if $n < m$. It leaves no room empty.) Eventually, the management attempts to list *all* the ways in which infinitely many rooms of Hotel Cosmos can be occupied. This leads to the diagonal argument and the discovery that $2^{|\mathbb{N}|} > |\mathbb{N}|$, followed by the uncountability of $\mathbb{R}$, the existence of transcendental numbers, and a discussion of the continuum hypothesis.

Thus, by page 70, he has covered the most important concepts and theorems of set theory, in an entertaining, intuitive, yet mathematically accurate way. Vilenkin's stories are far from random episodes; they form a well-structured, complete and

balanced whole, in which each story builds on those before—the way mathematics should be, of course.

So far, the book has dealt with the classical set theory of Cantor. The next section is for me the highlight of the book—a thoroughly intuitive approach to the notorious "counterintuitive" curves. We meet Koch's snowflake curve, Hilbert's space-filling curve, the Sierpinski carpet, and my favourite, a Jordan curve with positive area. The first example of such a curve was presented by Osgood in 1903, in a six-page paper in the *Transactions of the American Mathematical Society*. Vilenkin describes essentially the same example in just one page (including picture), which is breathtakingly simple. Having once attempted to read Osgood's paper, I cannot praise Vilenkin's achievement highly enough: what was once the ultimate affront to intuition has become intuitively obvious. The "crisis in intuition" seems to be over.

However, it would be foolish to be as dogmatically pro-intuition as Hans Hahn was anti. There is at least one counterintuitive cloud on the horizon: The axiom of choice. Vilenkin calls it "the baffling axiom" and mentions its most paradoxical consequence, the Banach-Tarski theorem about decomposing the ball and reassembling it into two, each the same size as the original. He also mentions that Gödel and Cohen proved its independence from the standard axioms of set theory. However, the importance of the axiom of choice elsewhere in mathematics, even elsewhere in the book, is not made very clear. There are opportunities to do so. Earlier, Vilenkin has introduced the concept of measurable set, so it would be worth mentioning that the Banach-Tarski paradox shows the limitations of this concept—it introduces *non*measurable sets. Likewise, it should be mentioned that the axiom of choice is needed to prove one of the first theorems in the book (page 57)—that countable sets are the smallest infinite sets.

Thus there are a few loose ends, but it is hard to complain about a mathematics book that leaves you eager to know more. In fewer than 150 pages Vilenkin has packed a whole course on infinity: its history, philosophy, and mathematical theory. Abe Shenitzer's translation is very smooth and natural, so one can read the whole book in an evening—and wish that it were longer. I cannot imagine a better introduction to set theory for beginners. They will surely hunger for a more advanced book, and Sierpinski's *Cardinal and Ordinal Numbers* might be a good next step.

For mathematicians, and teachers of mathematics, Vilenkin's most important message is that intuition has a vital role in set theory and analysis. Intuition can be *educated*, just like the logical faculty, to cope with infinity. The books on fractals have tried to tell us this, but Vilenkin has done it better, by getting to the mathematical heart of the problem. Understanding infinite geometric processes demands an understanding of infinite sets. It's obvious really, but Vilenkin makes this understanding easy and pleasant to acquire, and appropriately geometric. When fractals have gone out of style, and chaos is no longer the flavour of the month, some kind of "geometric set theory" or "set-theoretic geometry" may very well remain, and *In Search of Infinity* may still be the best introduction to it.

Department of Mathematics
Monash University
Clayton, Victoria 3168
Australia

Ramanujan: Letters and Commentary, by Bruce C. Berndt and Robert A. Rankin, American Mathematical Society, London Mathematical Society, History of Mathematics Series, Vol. 9 (1995), 347 pp., $79.00 (hc), $44.00 (pb).

Reviewed by **Krishnaswami Alladi**

Srinivasa Ramanujan is a truly exceptional figure in mathematical history. Born in a poor and orthodox Hindu family in rural India in 1887, Ramanujan did not even have a college degree. Yet, he made startling mathematical discoveries that challenged the finest minds of his and succeeding generations. Ramanujan communicated some of his remarkable findings in letters to G. H. Hardy of Cambridge University, who was so impressed that he helped to make arrangements for Ramanujan to go to England. During the brief period of five years (1914–19) that he spent in England, Ramanujan wrote several fundamental papers, some in collaboration with Hardy, which revolutionized certain areas of mathematics. His work was considered so original and profound that he was made Fellow of The Royal Society (FRS) and Fellow of Trinity College. Unfortunately, the rigors of life in England during World War I combined with his own peculiar habits contributed to a decline in his health. This forced him to return to India in 1919, where he died a year later. Despite his early death, the papers that he wrote in England and in India and the mass of unpublished material that he left behind in the form of two notebooks and loose sheets of paper, have had a deep and lasting impact and continue to inspire researchers today.

Ramanujan's story is sad, yet awe-inspiring in that someone who was so poor financially, who did not have a college education, and who grew up in such an old fashioned society, defied all odds and reached the pinnacle in mathematics, the most abstract and rigorous of all disciplines. Much has been written about Ramanujan and justifiably so, for he and his work are worthwhile studying from many points of view. For the layman, there is the wonderful book by Kanigel [3] describing Ramanujan's incredible life story. For those with a mathematical background, there is an account of Ramanujan's work by Hardy [2] in the form of twelve lectures, considered a classic in exposition. Then there are the Collected Papers [4], where at the beginning there are two charming biographies of Ramanujan written in contrasting styles, one by Hardy and another by P. V. Seshu Aiyar and R. Ramachandra Rao. For those who wish to delve directly into Ramanujan's identities and drink delight of his discoveries, there are the photostat versions of his two Notebooks [5] and The Lost Notebook [6] as well. And to top this all, there is a series of five volumes by Bruce Berndt [1] that thoroughly discuss the hundreds of "entries" made by Ramanujan in his famous notebooks, not to mention many research papers written in this century dealing with Ramanujan's work, in particular by George Andrews. So what could be added of great significance to this already impressive collection? As Shakespeare remarked, would it be like gilding refined gold, painting the lily, or adding another hue to the rainbow? Not at all, as this book under review demonstrates.

By assembling letters to, from, and about Ramanujan, and by giving an authoritative commentary on these letters, Berndt and Rankin have produced a book that should appeal to everyone with an interest in mathematics. Many of the letters contain mathematics. For these letters, Berndt and Rankin explain the results and

their significance with appropriate references. There are also letters from persons who knew Ramanujan and had an impact on his life. For such letters there are explanations of the connections these persons had with Ramanujan or his work. By reading the letters along with the comments, we get a better understanding of Ramanujan's personality, his life, and his remarkable work.

This book opens with a series of letters that describe how Ramanujan got a job at the Madras Port Trust and the efforts made by various well wishers to help him pursue his mathematical investigations. Although Ramanujan held a scholarship in his school, owing to his excessive preoccupation with mathematics, he neglected other subjects and consequently failed his college examinations in all subjects except mathematics. Naturally, without a college degree, he could not secure a job to support himself so that he could continue to do mathematics unhindered by financial difficulties. His family was poor and in no position to support him financially. Therefore Ramanujan approached several persons for help and showed them some of the remarkable identities recorded in his notebooks. These persons included Indians in well-placed positions as well as British administrators and professors in educational institutions in the city of Madras. While everyone of them was convinced of Ramanujan's unusual abilities and creativity, no one was able to judge the value of Ramanujan's work or understand it in the proper mathematical framework. One gentleman, R. Ramanchandra Rao, recognized that Ramanujan was doing very original work and gave him some financial support. It was clear that Ramanujan ought to come into contact with first-rate research mathematicians. With this favorable intention in mind, C. L. T. Griffith, Professor of Civil Engineering in Madras, wrote to M. J. M. Hill of the University College, London, about some of Ramanujan's work. Hill realized, as others had, that Ramanujan was very gifted, but when he saw Ramanujan's outrageous claims that

$$1 + 2 + 3 + \cdots = \frac{-1}{12}, \tag{1}$$

and

$$1^2 + 2^2 + 3^2 + \cdots = 0, \tag{2}$$

he felt that Ramanujan had bungled. And in a reply to Griffith, Hill expressed the opinion that these are the kind of blunders one would make without a formal mathematical training. After all, Abel has warned us that "divergent series are in general deadly, and anyone who dares to base a proof on them is doomed to failure". We know now that Ramanujan had not bungled here. On the contrary, these are significant assertions if viewed as follows. Consider the Riemann zeta function, which is defined by the series

$$\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s}, \tag{3}$$

for $\mathrm{Re}(s) > 1$. It is known, and this is a very significant fact, that $\zeta(s)$ admits an analytic continuation as a meromorphic function throughout the complex plane having $s = 1$ as its only singularity. Using this analytic continuation one can show that

$$\zeta(-1) = \frac{-1}{12} \quad \text{and} \quad \zeta(-2) = 0. \tag{4}$$

Observe that by setting $s = -1$ and $s = -2$ formally in (3) and using the values given in (4), the two claims made by Ramanujan in (1) and (2) follow.

But Ramanujan did not explain his claims in this fashion. In fact, he gave no explanation at all and that is what stunned Hill. We know now that Ramanujan had his own theory of infinite series. In particular, to each series, whether it is convergent or not, he associated a "constant". The constant he associated with the series in (1) was $-1/12$ and the constant he assigned the series in (2) was 0. Although no one at that time had any understanding of Ramanujan's methods, it was clear that he was unusually talented. So he deserved to have a scholarship or other financial support to enable him to continue his investigations. Since Ramanujan did not have a college degree, arranging a university scholarship was next to impossible, and so a job was initially secured for him in March, 1912 as a clerk in the Accounts Department at the Port Trust in Madras with the help of its Manager, S. Narayana Aiyar. The Chairman of the Port Trust, Sir Francis Spring, and Narayana Aiyar took a keen interest in Ramanujan's work and later helped him to secure a scholarship at the Madras University and also to prepare for his trip to England. In fact, Narayana Aiyar even worked with Ramanujan on some mathematical problems.

On February 13, 1913, Ramanujan wrote a letter to G. H. Hardy of Cambridge University. This letter is one of the most important and exciting mathematical letters ever written! After begging to introduce himself as a poor clerk in the Port Trust of Madras, Ramanujan gives a collection of amazing results he had obtained. These included representations in the form of series and integrals for the number of primes up to a given magnitude and related arithmetical functions, identities for beta and gamma functions, infinite series identities having arithmetical significance as well as consequences in the theory of elliptic and theta functions, and some unbelievable continued fraction evaluations in terms of certain algebraic numbers. There were some results in the letter that were well known and some that were wrong. But then, there were also many that were startlingly new and very deep. Hardy could prove a few of these, but there were others that defeated him completely. And Ramanujan had not given any hints as to how he went about proving them!

Hardy showed the letter to his colleague J. E. Littlewood and the two came to the conclusion that Ramanujan was a mathematician of the highest caliber and a genius on par with Euler and Jacobi in manipulative ability! Hardy was very excited about this and soon the news spread through Cambridge that at least another Jacobi in the making had been found. There is a very nice letter from Bertrand Russell to his sweetheart Lady Ottoline Morrell that is reproduced in this book wherein Russell says he "found Hardy and Littlewood in a state of wild excitement, because they believe they have discovered a second Newton." In replying to Ramanujan, Hardy pointed out the results that were wrong, those that were well known, and those that were new and most impressive. But he insisted that Ramanujan should supply proofs of his results. In his second letter to Hardy, which contained many more results, Ramanujan says the reason he did not describe his methods was that because they were so unusual, persons with formal training may not appreciate these unconventional approaches (as had been his experience with Hill). But Ramanujan was confident in the validity of his methods because in his first letter to Hardy he actually said—"the local mathematicians are unable to understand me in my higher flights". Both letters of Ramanujan to Hardy are reproduced in full with comments on their mathematical contents.

Hardy was convinced that Ramanujan should not waste any more time in India, but should come to England where his untutored genius would develop its full potential and be given a proper sense of direction. Thus Hardy urged Ramanujan

to come to England. Ramanujan initially was reluctant to go abroad owing to his religious background and pressure from his family, but he eventually agreed to go for the sake of mathematics.

During the five years that Ramanujan spent in England he wrote several papers that revealed the range and power of his mathematics. In collaboration with Hardy he wrote two great papers. In the first one, which appeared in 1917, he studied the most commonly occurring values of the number of prime factors of an integer. What is surprising here is that, although prime numbers have been studied since Greek antiquity and various arithmetical functions have been investigated in the subsequent centuries, it was the first systematic discussion of the number of prime divisors of an integer. This paper led to the creation of Probabilistic Number Theory. In another joint paper with Hardy, published in 1918, he gave an asymptotic formula for $p(n)$, the number of partitions of an integer n, using the *circle method*. This powerful analytic technique is now the standard tool in Additive Number Theory and its genesis may be traced to a formula that Ramanujan gave for the coefficients $c(n)$ of a certain series in his first letter to Hardy. What Hardy and Ramanujan produced was a series representation for $p(n)$ with terms involving the exponential function. Summing the series up to a certain number of terms dependent on n yielded a value whose nearest integer was $p(n)$. Later, Hans Rademacher made the important observation that replacing the exponential function by suitable hyperbolic functions converts the Hardy-Ramanujan representation into an infinite series that converges to the value $p(n)$. Subsequently, D. H. Lehmer showed that the Hardy-Ramanujan representation in terms of the exponential function is actually divergent as an infinite series. The correspondence between Hardy and Lehmer concerning the partition function is included in this book. It is worth noting that the formula Ramanujan communicated Hardy concerning $c(n)$ made use of hyperbolic functions, as in Rademacher's formula.

In England, Ramanujan also wrote many fundamental papers by himself. In one paper he established unexpected congruence properties for the partition function. After all, partitions represent an additive process and so it is surprising that certain divisibility (congruence) properties are valid here. In yet another famous paper on Modular Equations and Approximations to π, he gave several astonishing series representations, which are being used in this modern era of computers to calculate the digits of π. This paper also provided new insights into the properties of elliptic and modular functions.

Although his mathematical productivity was great, Ramanujan never got adjusted to the British way of life. The English winters were too harsh for him and he did not know how ,to protect himself adequately from the cold. He was a strict vegetarian and so did not want to eat the food served in the dining hall at Trinity College, preferring to cook his own food. But he was not very good at this. In particular, he did not eat a balanced diet. Things were made worse by the wartime rationing in England. All this had a catastrophic effect on his health. He was in and out of several sanatoria and hospitals because he was suspected to have tuberculosis. Even today we are not sure what exactly was Ramanujan's ailment. There is good reason to believe that he was suffering from hepatic amoebiasis, a parasitic infection of the liver or intestines that is found in tropical countries. There are many letters in this book that deal with Ramanujan's health problems.

Because Ramanujan's work was so original and important, Hardy felt that the Indian genius deserved to be made Fellow of The Royal Society and Fellow of Trinity. With Ramanujan's health declining rapidly, Hardy wanted these honors to

be conferred without delay. Hardy also felt that this would boost Ramanujan's spirit and have a positive effect on his health. So Hardy moved heaven and earth and finally succeeded in having these honors conferred on Ramanujan. This book contains a wonderful collection of letters between Ramanujan and his relatives and friends in India, wherein Ramanujan describes his mathematical successes in spite of the difficulties of living in England. In these letters he describes the particular Indian vegetarian dishes he is able to prepare. By reading the commentary of Berndt and Rankin, one gets a crash course on the curries and spices of South India! The correspondence between Ramanujan and Hardy while Ramanujan was in nursing homes is also included. It is amazing to see the quality of mathematics that Ramanujan communicated to Hardy from hospital beds in England! Finally, there are letters between Hardy and his British contemporaries giving us an idea of the efforts that went in to get Ramanujan elected Fellow of The Royal Society and Fellow of Trinity.

But Ramanujan's health did not improve substantially as Hardy and others had hoped. As soon as he was well enough to travel, Ramanujan returned to India in 1919. In India, his health worsened, and he died on April 26, 1920. But his mathematical powers had not diminished even in his final moments. He wrote one last letter to Hardy in which he described his latest discovery, the mock-theta functions, and expressed the opinion that these enter mathematics more naturally and beautifully than the false-theta functions of L. J. Rogers. These are now considered to be among Ramanujan's deepest contributions.

After Ramanujan's death, Hardy received the loose sheets of paper on which Ramanujan had scribbled mathematical formulae in his dying moments. We should be grateful to Janaki Ammal, Ramanujan's widow, for not throwing away his final jottings. Later Hardy handed these sheets to G. N. Watson at Birmingham. Contained in these sheets were several deep formulae for the mock-theta functions. Watson wrote two papers on the mock-theta functions but there was much more in these loose sheets that needed to be analyzed. After Watson's death, all the mathematical papers in his house were collected and deposited at the Wren Library in Cambridge University. Included in this collection were the last writings of Ramanujan. Surprisingly, the mathematical world remained unaware of the existence of these notes of Ramanujan until George Andrews accidentally came across them in 1976 while looking through the Watson collection at the Wren Library. The fascinating story of the discovery of Ramanujan's Lost Notebook by Andrews has been told many times. During the Ramanujan Centenary Celebrations in Madras, India, the Lost Notebook and other unpublished papers of Ramanujan as well as some correspondence between Hardy, Watson, and others after Ramanujan's death was brought out in a printed form **[6]**. These letters along with commentaries are included in this book.

At the beginning of this century Ramanujan was perceived by Hardy as a genius of the first magnitude who, without a sense of direction, had unfortunately wasted some of his best years rediscovering past work in India. Of course, Hardy did acknowledge that there was a significant number of new results in Ramanujan's work and that there were many that were so singularly original that he ranked among the great mathematicians in history. Today we are in a much better position to comprehend the grandeur and significance of Ramanujan's contributions. Hardy said that Ramanujan should have perhaps been born a century earlier during the great days of formulae. But Richard Askey points out that currently in physics there are incredible formulae in several variables that are being analyzed and a

genius like Ramanujan would be of invaluable help. As Askey puts it, "the great age of formulae may be over, but the age of great formulae is not!"

Speaking about Ramanujan, Hardy described him as "a poor, uneducated Hindu pitted against the accumulated wisdom of Europe". At this point I would like to compare Ramanujan with another equally astonishing figure, also from India, not in mathematics but in politics, namely, Mahatma Gandhi. Described by Sir Winston Churchill as "the half-naked fakir", Gandhi, wearing only a home spun loin cloth, stemmed the tide of British imperialism by his simplicity, honesty and adherence to non-violence. In a political world of charisma and diplomacy, who would have expected Gandhi to be a success? But Gandhian philosophy has had a lasting impact just as Ramanujan's equations scribbled on a piece of slate or on loose sheets of paper influence the work of some of the most sophisticated mathematicians today. Those turning to Ramanujan for inspiration would find in this book a fine introduction to the type of mathematics that Ramanujan did and a good sample of his great discoveries with references to over three hundred research papers, books, and articles. Those eager to learn about Ramanujan's life would get as a bonus, a glimpse into the culture and traditions of the Hindu way of life in British Colonial India. And finally, what better way to understand the man behind the mathematician Ramanujan than to read letters written by him and about him? Berndt, with the experience he has gained editing Ramanujan's notebooks, and Rankin, one of the veterans in this field, who knew Hardy, Littlewood, Watson, and other British contemporaries of Ramanujan, have combined perfectly to produce this book.

REFERENCES

1. B. C. Berndt, *Ramanujan's Notebooks*, Springer-Verlag, New York, Part I (1985), Part II (1989), Part III (1991), Part IV (1994), Part V (to appear).
2. G. H. Hardy, *Ramanujan*, Cambridge University Press, Cambridge (1940).
3. S. Kanigel, *The Man Who Knew Infinity*, Charles Scribners, New York (1991).
4. S. Ramanujan, *Collected Papers*, Cambridge University Press, Cambridge (1927).
5. S. Ramanujan, *Notebooks (2 Volumes)*, Tata Institute, Bombay (1957).
6. S. Ramanujan, *The Lost Notebook and Other Unpublished Papers*, Narosa, New Delhi (1987).

Mathematics Department
University of Florida
Gainesville, FL 32611

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of
Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *Collected Papers of Pierre Samuel, Volumes I–II.* Ed: Paulo Ribenboim. Papers in Pure & Appl. Math., V. 99. Queen's Univ, 1995. *Volume I*, xxix + 558 pp, (P), [ISBN 0-88911-693-8]; *Volume II*, vi + 531 pp, (P). [ISBN 0-88911-695-4]

Reference, P, L*. *Encyclopedia of Mathematics, Volumes 1–6.* Ed: M. Hazewinkel. Kluwer Academic, 1995, $699 set, [ISBN 1-55608-010-7 set]. *Volume 1: A-Integral–Coordinates*, ix + 919 pp; *Volume 2: Coproduct–Hausdorff-Young Inequalities*, 963 pp; *Volume 3: Heaps and Semi-Heaps–Moments, Method of (in Probability Theory)*, 950 pp; *Volume 4: Monge-Ampère Equation–Rings and Algebras*, 929 pp; *Volume 5: Risk of a Statistical Procedure–Zygmund Class of Functions*, 906 pp; *Volume 6: Subject Index–Author Index*, xiii + 732 pp. Unabridged paperback version of the 10-volume hardbound edition.

Recreational Mathematics. *Puzzles, Mazes and Numbers.* Charles Snape, Heather Scott. Cambridge Univ Pr, 1995, 128 pp. [ISBN 0-521-46500-1] Intended for anyone ≥ 10, but illustrations, format aimed at children. Puzzles are mainly elementary, but some are quite challenging. Topics: logic, geometry, patterns, knots. Some answers in back; no explanations. JCS

Finite Mathematics, T(13–14: 1–3). *College Mathematics: A Graphing Calculator Approach.* Ruric Wheeler, Karla Neal, Roseanne Hofmann. Wiley, 1996, xvi + 973 pp, $75.95. [ISBN 0-471-05720-7] Mathematical topics: matrix arithmetic and linear programming, finance math, elementary probability and statistics, calculus. Includes exponential and logarithmic functions, but not trigonometric functions. Relies heavily on graphing calculator to develop intuition, make estimates, verify results, check reasonableness of solutions. Emphasizes conceptual understanding more than rote application of techniques. Encourages intelligent use of calculator. LB

Logic, T(15–16: 1, 2), L. *The Incompleteness Phenomenon: A New Course in Mathematical Logic.* Martin Goldstern, Haim Judah. AK Peters, 1995, xiii + 247 pp, $49.95. [ISBN 1-56881-029-6] First two chapters cover first-order logic. Third chapter introduces tools of model theory: elementary substructures, ultraproducts, types. Fourth chapter provides proof of simplified form of Gödel's incompleteness theorem. Numerous exercises. KES

Logic, P. *Logic Colloquium '92.* Eds: László Csirmaz, Dov M. Gabbay, Maarten de Rijke. Stud. in Logic, Lang. & Inform. Center for the Study of Language & Information (Ventura Hall, Stanford U., Stanford, CA 94305), 1995, xiii + 319 pp, $22.95 (P); $49.95. [ISBN 1-881526-98-4; 1-881526-97-6] Invited and contributed papers from a 1992 conference in Veszprém, Hungary.

Logic, P. *Modal Logic and Process Algebra: A Bisimulation Perspective.* Eds: Alban Ponse, Maarten de Rijke, Yde Venema. CSLI Lect. Notes, No 53. Center for the Study of Language & Information (Ventura Hall, Stanford U., Stanford, CA 94305), 1995, xix + 327 pp, $22.95 (P); $49.95. [ISBN 1-881526-96-8; 1-

881526-95-X] 14 papers from a 1994 workshop "Three Days of Bisimulation" at the Centre for Mathematics and Computer Sciences in Amsterdam.

Combinatorics, P. *Combinatorics Advances.* Eds: Charles J. Colbourn, Ebadollah S. Mahmoodian. Math. & Its Applic., V. 329. Kluwer Academic, 1995, xv + 324 pp, $170. [ISBN 0-7923-3574-0] Papers from the 25th Annual Iranian Mathematics Conference held in March 1994 at Sharif University of Technology in Tehran.

Combinatorics, P. *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms.* ACM & SIAM, 1996, 586 pp, $62 (P). [ISBN 0-89871-366-8] 65 papers presented in January 1996.

Number Theory, P. *Selected Papers on Number Theory and Algebraic Geometry.* Ed: Katsumi Nomizu. AMS Transl. Ser. 2, V. 172. AMS, 1996, ix + 91 pp. [ISBN 0-8218-0445-6] 5 papers that originally appeared in the Mathematical Society of Japan's journal *Sugaku.*

Number Theory, P, L. *Handbook of Number Theory.* D.S. Mitrinovic, J. Sándor, B. Crstici. Math. & Its Applic., V. 351. Kluwer Academic, 1996, xxvi + 622 pp, $277. [ISBN 0-7923-3823-5] "Most important results from some parts of number theory, which are expressed by inequalities or by estimates." "Focuses on the most important arithmetic functions." BH

Group Theory, P. *The Classical and Quantum 6j-symbols.* J. Scott Carter, Daniel E. Flath, Masahico Saito. Math. Notes, V. 43. Princeton Univ Pr, 1995, ix + 164 pp, $22.50 (P). [ISBN 0-691-02730-7]

Group Theory, P. *Spectral Decomposition and Eisenstein Series.* C. Moeglin, J.-L. Waldspurger. Tracts in Math., V. 113. Cambridge Univ Pr, 1995, xvii + 338 pp, $80. [ISBN 0-521-41893-3]

Algebra, P. *The Theory of Lattice-Ordered Groups.* V.M. Kopytov, N. Ya. Medvedev. Math. & Its Applic., V. 307. Kluwer Academic, 1994, xvi + 400 pp, $176. [ISBN 0-7923-3169-9] Introduction to partially-ordered groups (emphasizing lattice orders) and ordered permutation groups. Surveys recent work, including quasi-varieties of lattice-ordered groups. TH

Algebra, P. *Abelian Groups and Modules.* Eds: Alberto Facchini, Claudia Menini. Math. & Its Applic., V. 343. Kluwer Academic, 1995, xx + 517 pp, $229. [ISBN 0-7923-3756-5] Papers from a 1994 conference at the University of Padova.

Algebra, T(18), P. *An Introduction to Homological Algebra.* Charles A. Weibel. Stud. in Adv. Math., V. 38. Cambridge Univ Pr, 1995, xiv + 450 pp, $29.95 (P); $49.95. [ISBN 0-521-55987-1; 0-521-43500-5] Aims to make the tools of homological algebra accessible to nonexperts. Comprehensive, up-to-date account. TH

Algebra, P. *Introduction to the Theory of Topological Rings and Modules.* V.I. Arnautov, S.T. Glavatsky, A.V. Mikhalev. Pure & Appl. Math., V. 197. Marcel Dekker, 1996, v + 502 pp, $165. [ISBN 0-8247-9323-4] General theory of topological rings, including completion, semi-direct products, and non-discrete topologizations. Contains a list of open problems, and an extensive bibliography. TR

Algebra, P. *Categorical Structure of Closure Operators With Applications to Topology, Algebra and Discrete Mathematics.* D. Dikranjan, W. Thołlen. Math. & Its Applic., V. 346. Kluwer Academic, 1995, xvii + 358 pp, $169. [ISBN 0-7923-3772-7]

Algebra, T(17: 2). *Algebraic Structures.* George R. Kempf. Friedr Vieweg & Sohn, 1995, ix + 165 pp, $29 (P). [ISBN 3-528-06583-4] Concise but comprehensive introduction to algebra. Basic material on groups, rings, fields, and modules as well as modern linear algebra (tensor products), quadratic and alternating forms, Noetherian rings and localization, Dedekind domains, etc. Exercises interspersed throughout text. LC

Calculus, T(13: 2). *Calculus for Business, Economics, and the Social and Life Sciences, Sixth Edition.* Laurence D. Hoffmann, Gerald L. Bradley. McGraw-Hill, 1996, xviii + 844 pp. [ISBN 0-07-029372-4] New edition (*Fifth Edition*, TR, February 1993) adds writing, essay, and graphing calculator problems. Additional emphasis on the sciences and more detailed business/economics applications. Improved treatment of curve-sketching, streamlined treatment of integration, and revised treatment of series and trigonometric functions. JCS

Calculus, P*, L*. *Calculus: The Dynamics of Change.* A. Wayne Roberts. MAA Notes No. 39. MAA, 1996, x + 166 pp, $34.95 (P). [ISBN 0-88385-098-2] A ten-year retrospective on the national movement to make calculus "a pump not a filter." Diverse, brief essays offer commentary on visions and goals; planning and change; assessment and final exams; and connections pre- and post-calculus. Sample final exams invite comparison with the similar collection published in *Calculus for a New Century* (TR, February 1988), but no comparative anal-

ysis is provided. Contains a wealth of useful insight to aid would-be reformers, save compelling evidence to quiet critics. LAS

Real Analysis, T?(14), S. *Elementary Analysis through Examples and Exercises.* John Schmeelk, Djurdjica Takači, Arpad Takači. Texts in Math. Sci., V. 10. Kluwer Academic, 1995, ix + 319 pp, $156. [ISBN 0-7923-3597-X] Non-traditional presentation. Develops theory with minimum of discourse followed by worked-out examples and exercises for the reader. Many good problems, some non-trivial. BH

Complex Analysis, T*(18), P*, L. *Potential Theory in the Complex Plane.* Thomas Ransford. London Math. Soc. Stud. Texts, V. 28. Cambridge Univ Pr, 1995, x + 232 pp. [ISBN 0-521-46120-0] Nicely motivated, clear exposition of the basic theory. Final, more advanced chapter deals with applications in functional analysis, approximation theory, and dynamical systems. BH

Complex Analysis, T(13: 1), C, L. *Complex Variables and Applications, Sixth Edition.* James Ward Brown, Ruel V. Churchill. McGraw-Hill, 1996, xvi + 386 pp, $49.50, with IBM disk. [ISBN 0-07-912147-0] Changes from the *Fifth Edition* include new chapter on applications of residues, primarily dealing with inverse Laplace transforms; more emphasis on indented contours; improved proofs of Taylor's and Laurent's theorems; revision of power series development. Comes with disk containing abbreviated version of $f(z)$ software. BH

Complex Analysis, P. *Modern Methods in Complex Analysis.* Eds: Thomas Bloom, *et al.* Annals of Math. Stud., No. 137. Princeton Univ Pr, 1995, xiv + 342 pp, $35 (P). [ISBN 0-691-04428-7] 15 papers from a March 1992 conference at Princeton University honoring R.C. Gunning and J.J. Kohn.

Differential Equations, P. *Rigid Local Systems.* Nicholas M. Katz. Annals of Math. Stud., No. 139. Princeton Univ Pr, 1996, vii + 223 pp, $22.50 (P); $49.50. [ISBN 0-691-01118-4; 0-691-01119-2]

Partial Differential Equations, P. *Analytic Semigroups and Semilinear Initial Boundary Value Problems.* Kazuaki Taira. London Math. Soc. Lect. Note Ser., V. 223. Cambridge Univ Pr, 1995, x + 164 pp, $32.95 (P). [ISBN 0-521-55603-1]

Numerical Analysis, T(15–16: 1), L. *Numerical Methods Using MATLAB.* John Penny, George Lindfield. Ellis Horwood, 1995, xii + 328 pp, $35 (P). [ISBN 013-030966-4] Describes the MATLAB computer package and develops the usual algorithms found in an introductory numerical analysis text. A few short exercises appear at the end of each chapter; answers in the back. SM

Numerical Analysis, P. *Numerical Methods for the Solution of Ill-Posed Problems.* A.N. Tikhonov, *et al.* Math. & Its Applic., V. 328. Kluwer Academic, 1995, ix + 253 pp, $137. [ISBN 0-7923-3583-X]

Operator Theory, P. *The Heat Kernel Lefschetz Fixed Point Formula for the Spin-c Dirac Operator.* J.J. Duistermaat. Progress in Nonlinear Diff. Eqns. & Their Applic., V. 18. Birkhäuser Boston, 1996, viii + 247 pp, $38.50. [ISBN 0-8176-3865-2]

Operator Theory, T(18), S. *Spectral Theory and Differential Operators.* E.B. Davies. Stud. in Adv. Math., V. 42. Cambridge Univ Pr, 1995, ix + 182 pp, $49.95. [ISBN 0-521-47250-4] The spectral theorem may be paraphrased by saying that any self-adjoint linear operator is unitarily equivalent to a multiplication operator. Not surprisingly, this result is quite useful in deciphering the structure of differential operators. Emphasizes second-order elliptic differential operators, concluding with a chapter on Schrödinger operators. SA

Operator Theory, P. *Operator Theory and Boundary Eigenvalue Problems.* Eds: I. Gohberg, H. Langer. Operator Theory: Adv. & Applic., V. 80. Birkhäuser Boston, 1995, x + 313 pp, $123. [ISBN 0-8176-5275-2] Proceedings of a July 1993 workshop in Vienna, Austria.

Operator Theory, P. *Complexes of Differential Operators.* Nikolai N. Tarkhanov. Math. & Its Applic., V. 340. Kluwer Academic, 1995, xviii + 396 pp, $198. [ISBN 0-7923-3706-9] In the author's opinion, "the universe around us seems to be overdetermined." Complexes of differential operators arise naturally as descriptions of overdetermined systems of differential equations. This book is a systematic study of these so-called differential complexes. SA

Functional Analysis, P. *Functional Analysis in China.* Eds: Bingren Li, *et al.* Math. & Its Applic., V. 356. Kluwer Academic, 1996, xv + 373 pp, $182. [ISBN 0-7923-3880-4] 30 survey and research papers summarize work in China since the 1940's.

Functional Analysis, T(16–17), P. *Theory of Distributions: A Non-technical Introduction.* Ian Richards, Heekyung Youn. Cambridge Univ Pr, 1995, ix + 147 pp, $49.95; $19.95 (P). [ISBN 0-521-37149-X; 0-521-

55890-5] Treats much of the theory of distributions, or "generalized functions," without advanced methods (measure theory, topological vector spaces). Contains most of what a nonspecialist needs to know—Fourier transforms, tempered distributions, multiplication and convolution for distributions. Clear and elementary presentation. (First published 1990, TR, February 1991.) SA

Functional Analysis, P. *Families of Exponentials: The Method of Moments in Controllability Problems for Distributed Parameter Systems.* Sergei A. Avdonin, Sergei A. Ivanov. Cambridge Univ Pr, 1995, xv + 302 pp, $54.95. [ISBN 0-521-45243-0]

Functional Analysis, T(18), S, P. *The Hilbert Transform of Schwartz Distributions and Applications.* J.N. Pandey. Pure & Appl. Math. Wiley, 1996, xvi + 262 pp, $64.95. [ISBN 0-471-03373-1] "Objective is to demonstrate the wide applicability of Hilbert transform techniques." Topics: the Riemann-Hilbert problem; applications in wing theory, theory of elasticity, spectroscopy, and high energy physics; Hilbert transforms of Schwarz distributions and periodic distributions. Assumes background in functional analysis. Some exercises. BH

Analysis, P. *Null-Additive Set Functions.* Endre Pap. Math. & Its Applic., V. 337. Kluwer Academic, 1995, xii + 315 pp, $154. [ISBN 0-7923-3658-5]

Analysis, P. *Potential Theory on Infinite-Dimensional Abelian Groups.* Alexander Bendikov. Stud. in Math., V. 21. Walter de Gruyter, 1995, vi + 184 pp, DM 168. [ISBN 3-11-014283-X]

Analysis, P, L. *Wavelets: An Analysis Tool.* M. Holschneider. Oxford Math. Mono. Clarendon Pr, 1995, xiii + 423 pp, $85. [ISBN 0-19-853481-7] "Introductory text" intended to "fill a gap between the book by Ingrid Daubechies and more advanced books by Yves Meyer and Ronald Coifman." "Focuses on the continuous wavelet transform and some of its applications." Topics include localization properties of wavelet transforms, theory of frames, construction of orthonormal wavelet bases, multi-resolution analysis, fractals. BH

Analysis, T(17), P, L. *Measure, Integration and Function Spaces.* Charles Swartz. World Scientific, 1994, x + 277 pp, $61. [ISBN 981-02-1610-6] Develops standard theory of Lebesgue integration, gives short introduction to functional analysis, and then studies various function spaces, including L^p spaces and spaces of finitely additive and countably additive functions. Strong emphasis on role of countable additivity in measure theory. Numerous exercises. BH

Topology, P. *Topological Methods in Differential Equations and Inclusions.* Eds: Andrzej Granas, Marlène Frigon. NATO ASI Ser. C, V. 472. Kluwer Academic, 1995, xv + 522 pp, $245. [ISBN 0-7923-3678-X] Papers from a 1994 NATO Advanced Study Institute at the Université de Montrëal.

Topology, P. *Prospects in Topology: Proceedings of a Conference in Honor of William Browder.* Ed: Frank Quinn. Annals of Math. Stud., No. 138. Princeton Univ Pr, 1995, xiii + 340 pp, $29.95 (P); $65. [ISBN 0-691-02728-5; 0-691-02729-3] 18 papers from a March 1994 conference at Princeton University.

Topology, P. *The Seiberg-Witten Equations and Applications to the Topology of Smooth Four-Manifolds.* John W. Morgan. Math. Notes, V. 44. Princeton Univ Pr, 1996, vi + 128 pp, $19.95 (P). [ISBN 0-691-02597-5]

Operations Research, P. *Scheduling Theory and its Applications.* Eds: Philippe Chrétienne. Wiley, 1995, xv + 365 pp, $69.96. [ISBN 0-471-94059-3] Tutorials, surveys, and research articles from a 1992 meeting in France.

Optimal Control, P. *Lecture Notes in Control and Information Sciences-209: Robust Control of Infinite Dimensional Systems, Frequency Domain Methods.* C. Foias, H. Özbay, A. Tannenbaum. Springer-Verlag, 1996, 218 pp, $54 (P). [ISBN 3-540-19994-2]

Probability, T(18), S. *Probability Measures on Semigroups: Convolution Products, Random Walks, and Random Matrices.* Göran Högnäs, Arunava Mukherjea. Univ. Ser. in Math. Plenum Pr, 1995, xii + 388 pp, $89.50. [ISBN 0-306-44964-1] Up-to-date information on theory of random walks with values in semigroups. (For example, a Markov chain can be regarded as a random walk on the transformation of the state space.) Familiarity with probability, measure theory, and group theory required. SA

Stochastic Processes, T(18: 12), S. *Stochastic Processes: General Theory.* M.M. Rao. Math. & Its Applic., V. 342. Kluwer Academic, 1995, xii + 623 pp, $259. [ISBN 0-7923-3725-5] Revision of the 1979 *Stochastic Processes and Integration* (TR, April 1982). A serious, careful exposition. Measure theory-based, it covers existence questions, stochastic integration, and differential equations with emphasis on Martingale techniques. Many exercises. Large and useful bibliography. TAV

Stochastic Processes, T(12–17: 2), L*.** *Stochastic Processes, Second Edition.* Sheldon M. Ross. Wiley, 1996, xv + 510 pp, $87.95. [ISBN 0-471-12062-6] Revision of a fine text (TR, April 1984); each chapter has been fine-tuned. Does not assume measure theory. Developments of Poisson and related processes, and Markov processes are particularly clear and instructive. Very appealing as a text for a year-long course. Many well-chosen problems—some with answers. TAV

Elementary Statistics, T(13: 1). *Statistics: Principles and Methods, Third Edition.* Richard A. Johnson, Gouri K. Bhattacharyya. Wiley, 1996, xv + 720 pp, $70.95. [ISBN 0-471-04194-7] Maintains goals of *Second Edition* (TR, January 1993): developing an understanding of how findings from sample data can be extended to a population, familiarity with basic statistical concepts, and an appreciation of the computer in data analysis. Many examples from real life settings. Minitab output incorporated throughout. Improvements over previous editions include more emphasis on understanding data, more data-based exercises, and a more thorough coverage of regression analysis. Electronic data bank available on disk. RS

Elementary Statistics, T(13–14: 1, 2), C. *Introductory Statistics.* Sheldon M. Ross. McGraw-Hill, 1996, xviii + 807 pp. [ISBN 0-07-912244-2] Readable, engaging text with a wealth of good examples and exercises using real data. Accompanied by disk containing larger data sets. Each chapter concludes with a Minitab lab. Includes material on ANOVA, multiple regression, nonparametric tests, quality control. Little discussion of sampling methods, experimental design or handling data from observational studies. LB

Mathematical Statistics, T(18: 1), P, L. *Multivariate Analysis, Part 2: Classification, Covariance Structures and Repeated Measurements.* W.J. Krzanowski, F.H.C. Marriott. Kendall's Lib. of Stat., V. 2. Arnold (Copub: Halsted Pr), 1995, viii + 280 pp, $54.95. [ISBN 0-470-23413-X] Second part of the first volume of this new series (Series and *Part 1*, TR, May 1995). Covers more specialized topics. Extensive references. RSK

Statistical Methods, T(14–17: 1, 2). *A Second Course in Statistics: Regression Analysis, Fifth Edition.* William Mendenhall, Terry Sincich. Prentice Hall, 1996, xxi + 899 pp. [ISBN 0-13-396821-9] Emphasizes model building and applications of regression analysis. Many worked examples and problems involving real data; five case study chapters. This edition features expanded range of applications and examples, more computer printouts (SAS, SPSS, Minitab, ASP), and an updated appendix on statistical software commands. RS

Statistical Methods, T*(15–17: 1–3), C. *Applied Linear Statistical Models, Fourth Edition.* John Neter, *et al.* Richard D Irwin, 1996, xv + 1408 pp, with disk. [ISBN 0-256-11736-5] Applications of linear statistical models in regression, analysis of variance, and experimental design. New edition features expanded coverage of regression diagnostics, bootstrapping, weighted least squares, generalized linear models, response surface methodology, and random and mixed ANOVA models. Chapters on nonlinear, logistic regression, and correlation analysis have been re-introduced. Diskette has data sets. (*Second Edition*, TR, January 1986.) RS

Statistical Methods, T(17–18), P. *Pattern Classification: A Unified View of Statistical and Neural Approaches.* Jürgen Schürmann. Wiley, 1996, xvii + 373 pp, $64.95. [ISBN 0-471-13534-8] Statistical decision theory approach and estimating functions using statistical modeling, multilayer perceptron regression, and radial basis functions. Extensive chapter on polynomial regression and chapters on measurement and feature selection, reject criteria and classifier performance, and combining classifiers. Simplified two-dimensional example used throughout to illustrate techniques and to stress similarities among methods. RS

Statistical Methods, T(16–17), C. *Elementary Survey Sampling, Fifth Edition.* Richard L. Scheaffer, William Mendenhall III, Lyman Ott. Duxbury Pr, 1996, x + 501 pp, $64.50, with disk. [ISBN 0-534-24342-8] Design, data collection, and analysis issues. Topics include stratified random, systematic, and cluster sampling; ratio, regression, and difference estimation. Emphasizes practical aspects including sample size determination. Appendix contains derivations of major results. Uses Minitab throughout. Minitab and SAS macros in appendix and on data disk. RS

Statistical Methods, T(18). *Introduction to Statistical Time Series, Second Edition.* Wayne A. Fuller. Ser. in Prob. & Stat. Wiley, 1996, xxii + 698 pp, $69.95. [ISBN 0-471-55239-9] Theorem-proof format. New edition includes Wold decomposition, partial autocorrelation, long memory processes, the Kalman filter, nonlinear estimation, estimated generalized least squares, and the incorporation of asymptotic theory. (*First Edition*, TR, May 1977.) RS

Statistics, P*. *Handbook of the Normal Distribution, Second Edition, Revised and Expanded.*

Jagdish K. Patel, Campbell B. Read. Stat.: Textbooks & Mono., V. 150. Marcel Dekker, 1996, ix + 431 pp, $135. [ISBN 0-8247-9342-0] Chapter on Wiener and Gaussian processes has been removed, bivariate normal material has been expanded to two chapters, and two chapters covering estimation procedures for normally distributed samples have been added. (*First Edition*, TR, November 1982.) RSK

Mathematical Computing, S(13–15). *Maple V: Learning Guide.* K.M. Heal, M.L. Hansen, K.M. Rickard. Springer-Verlag, 1996, ix + 269 pp, $24 (P). [ISBN 0-387-94536-9] First chapter includes tutorials explaining worksheet interface and on-line help system. Subsequent chapters use examples to illustrate numerical computations, graphics, evaluation and simplification, and input and output. One chapter on examples from calculus. KES

Computer Systems, P. *Porting UNIX Software: From Download to Debug.* Greg Lehey. O'Reilly & Associates, 1995, xix + 515 pp, $29.95 (P). [ISBN 1-56592-126-7]

Theory of Computation, P. *Bounded Arithmetic, Propositional Logic, and Complexity Theory.* Jan Krajíček. Ency. of Math. & Its Applic., V. 60. Cambridge Univ Pr, 1995, xiv + 343 pp, $59.95. [ISBN 0-521-45205-8] Bounded arithmetic theory is subtheory of Peano arithmetic whose models are nonnegative parts of discretely ordered commutative rings. One approach to study of $P \pm NP$ problem is to prove that this inequality holds in nonstandard models of systems of bounded arithmetic. Book presents basics of topics in title and explains relations between them. KES

Computer Science, T*(15–18: 1, 2), L. *Mathematical Methods in Artificial Intelligence.* Edward A. Bender. IEEE Computer Society Pr, 1996, xx + 636 pp, $55. [ISBN 0-8186-7200-5] Focus on mathematical tools for designing AI tools. Key topics: trees and search, predicate logic, use of first-order logic in real world problems, pattern classification. RJA

Computer Science, P. *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence.* Eds: Giulianella Coletti, Didier Dubois, Romano Scozzafava. Plenum Pr, 1995, ix + 308 pp, $89.50. [ISBN 0-306-45076-3] Invited and contributed papers plus the roundtable discussion from a workshop at the International School of Mathematics "G. Stampacchia," Erice, Italy, June 1994. RJA

Computer Science, T(16–18: 1), S, P. *Epistemic Logic for AI and Computer Science.* J.-J. Ch. Meyer, W. van der Hoek. Tracts in Theoret. Comp. Sci., V. 41. Cambridge Univ Pr, 1995, xiii + 354 pp, $54.95. [ISBN 0-521-46014-X] A logic of knowledge and belief based on modal logic that deals with necessity and contingency. Contains multi-agent epistemic logic, application to distributed systems, knowledge vs. ignorance, default and counterfactual reasoning. Exercises with answers. RJA

Computer Science, P. *Lecture Notes in Control and Information Sciences-210: Neural Network Models, An Analysis.* Philippe De Wilde. Springer-Verlag, 1996, 153 pp, $40 (P). [ISBN 3-540-19995-0]

Applications (Biological Science), T(15–16: 1), P, L. *Image Analysis for the Biological Sciences.* C.A. Glasbey, G.W. Horgan. Wiley, 1995, xi +218 pp, $49.95. [ISBN 0-471-93726-6] Begins with an excellent general discussion of image analysis and its importance. Emphasizes the analysis of small scale images produced in laboratory environments. Extensive bibliography and links to images available over the internet. SM

Applications (Engineering), P. *Lecture Notes in Control and Information Sciences-211: Balanced Control of Flexible Structures.* Wodek Gawronski. Springer-Verlag, 1996, xiv + 262 pp, $50 (P). [ISBN 3-540-76017-2]

Applications (Physics), P. *Symmetries in Science VIII.* Ed: Bruno Gruber. Plenum Pr, 1995, x + 465 pp, $125. [ISBN 0-306-45119-0] Papers from a 1994 symposium in Bregenz, Austria.

Applications (Physics), P. *Fractals and Disordered Systems, Second Revised and Enlarged Edition.* Armin Bunde, Shlomo Havlin. Springer-Verlag, 1996, xxi + 408 pp, $69.50. [ISBN 3-540-56219-2] Topics include percolation, fractal growth, fractures, transport across irregular membranes, and cellular automata. (*First Edition*, TR, May 1992.)

Applications (Statistical Mechanics), P. *Physics of Critical Fluctuations.* Yuli M. Ivanchenko, Alexander A. Lisyansky. Grad. Texts in Contemp. Physics. Springer-Verlag, 1995, xv + 390 pp, $59. [ISBN 0-387-94414-1]

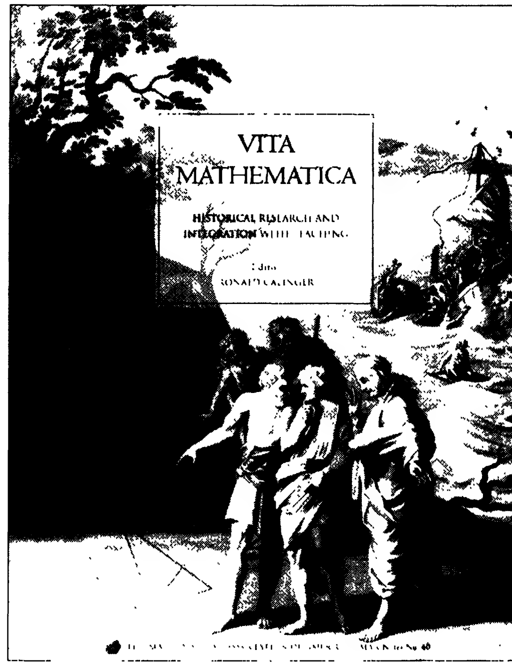
Reviewers

SA: Steve Abbott, St. Olaf; RJA: Richard J. Allen, St. Olaf; LB: Lynne Baur, Carleton; LC: Laura Chihara, St. Olaf; TH: Tom Halverson, Macalester; BH: Bruce Hanson, St. Olaf; RSK: Richard S. Kleber, St. Olaf; SM: Steve McKelvey, St. Olaf; TR: Tommy Ratliff, St. Olaf; RS: Richard Single, St. Olaf; JCS: Janice Sklensky, St. Olaf; KES: Kay E. Smith, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; TAV: Theodore A. Vessey, St. Olaf.

Vita Mathematica

Historical Research and Integration with Teaching

Ronald Calinger, Editor



The use of the history of mathematics in the teaching of mathematics at all levels is an idea whose time has come. To use history in the teaching of undergraduate mathematics, the instructor must be familiar with the history as well as the mathematics. *Vita Mathematica* will enable college teachers to learn the relevant history of various topics in the undergraduate curriculum and help them incorporate this history in their teaching.

For example, should calculus be approached from a geometric or an algebraic point of view? The book shows us how two important eighteenth century mathematicians, Colin Maclaurin and Joseph-Louis Lagrange, understood the calculus from these different standpoints and how their legacy is still important in teaching calculus today. We also learn why Lagrange's algebraic approach dominated teaching in Germany in the nineteenth century. Some of the reasons for this are related to the appropriate foundations of the calculus, and so the book traces the ancient history of one of the possible foundations, the concept of indivisibles. Even though we generally do not use this concept formally today, many ideas for a heuristic approach to the calculus can be developed out of his study.

Vita Mathematica contains numerous other articles dealing with calculus, with algebra, combinatorics, graph theory, and geometry, as well as more general articles on teaching courses for prospective teachers.

This volume, then, demonstrates that the history of mathematics is no longer tangential to the mathematics curriculum, but in fact deserves a central role.

Catalog Code: NTE40
350 pp., Paperbound, 1996, ISBN 0-88385-097-4
List: $34.95 MAA Member: $29.00

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	NTE40	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

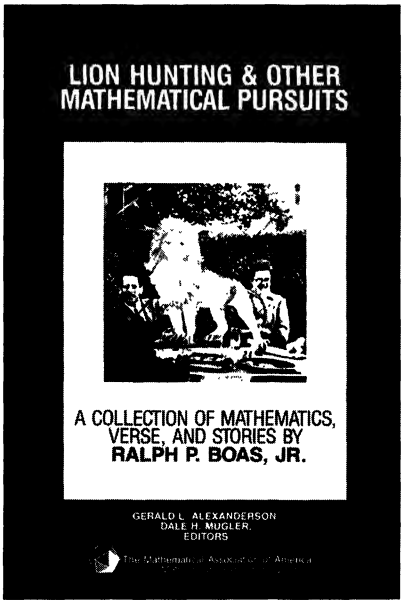
Signature ______________________

Lion Hunting

and Other Mathematical Pursuits

A Collection of Mathematics, Verse, and Stories
by Ralph P. Boas, Jr.

Gerald L. Alexanderson and
Dale H. Mugler, Editors



I highly recommend Lion Hunting and Other Mathematical Pursuits to high school mathematics clubs, mathematics teachers of all levels, and anyone interested in mathematics. Perhaps the most important features of this book is how it subtly makes the reader aware of the nature of mathematics.

– The Mathematics Teacher

As a young man at the Institute for Advanced Study in Princeton, Ralph Philip Boas, Jr., together with a group of other mathematicians, published a light-hearted article on the "mathematics of lion hunting" under a pseudonym (1938). This sparked a sequence of articles on the topic, several of which are drawn together in this book.

Lion Hunting includes an assortment of articles that show the many facets of this remarkable mathematician, editor, writer, and teacher. Along with a variety of his lighter mathematical papers, the collection includes Boas' verse and short stories, many of which are appearing for the first time. Anecdotes and recollections of his numerous experiences and of his work and meetings with many distinguished mathematicians and scientists of his day are also included as well as photographs taken by Boas of Hardy, Littlewood, Besicovitch, Weil, and others.

The mathematical articles in this collection cover a range of topics. They include articles on infinite series, the mean value theorem, indeterminate forms, complex variables, inverse functions, extremal problems for polynomials and more. A special section of this book is devoted to articles about the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we make mathematics intelligible?"

Boas's wit and playful humor are reflected in the verses included in this collection. The verses reflect the phases of his career as author, editor, teacher, department chair, and lover of literature. A section of the book describes the feud that Boas supposedly had with Bourbaki. Also included are many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995, ISBN 0-88385-323-X
List: $39.95 MAA Member: $28.95
Catalog Code: DOL-15

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ____________

Address ____________

City ____________

State ________ Zip ________

Qty.	Catalog Code	Price	Amount
____	DOL-15	____	____
____	____	____	____
		TOTAL	____

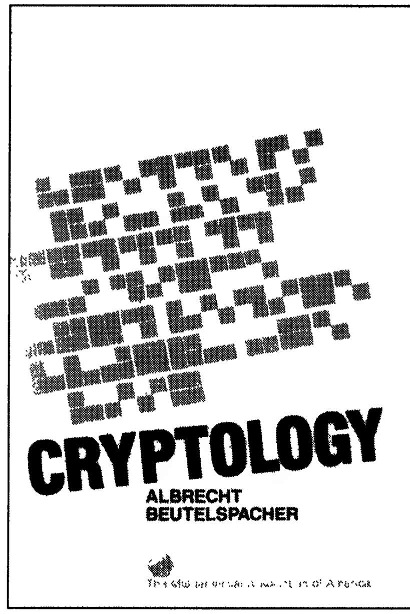
Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________ Expires ___/___

Signature ____________

CRYPTOLOGY

Albrecht Beutelspacher



This fascinating little book is eminently readable, and it is a great deal of fun to peruse... the book is a real treat. We need more books like this, crafted by expert hands yet crafted so that the general reader can enjoy them.
—Bulletin of The Institute of Combinatorics and Its Applications

This excellent and entertaining book is suitable for a first course in cryptology for mathematical enthusiasts. An abundance of exercises and an excellent list of related references are included.
—The Mathematics Teacher

In spite of the light-hearted style in which the book is written throughout, it is a serious—and successful—attempt to explain the basis of coding and decoding messages...I can strongly recommend this book to anyone who wants a brief but comprehensive, eminently readable, and up-to-date introduction to this increasingly popular topic.
— The Mathematical Gazette

All of cryptology is covered in this work...Occupying a niche in the halls of the ivory tower of pure mathematics for nearly two millennia, number theory now forms a pillar of modern society. This book is the best explanation available today of how that pillar was constructed.
— Charles Aschbacher

A model to follow in order to make mathematics better known and understood. Accessible to a broad audience. Have fun reading this book, while you are getting a better understanding of cryptology.
— Bulletin of the Belgian Mathematics Society

How can messages be transmitted secretly? How can one guarantee that the message arrives safely in the right hands exactly as it was transmitted? Cryptology—the art and science of "secret writing"—provides ideal methods to solve these problems of data security.

The book is fun to read, and the author presents the material clearly and simply. Many exercises and references accompany each chapter.

176 pp., Paperbound, 1994
ISBN 0-88385-504-6
List: $34.00
MAA Member: $26.95
Catalog Code: CRYPT/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (202) 265-2384

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	CRYPT/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

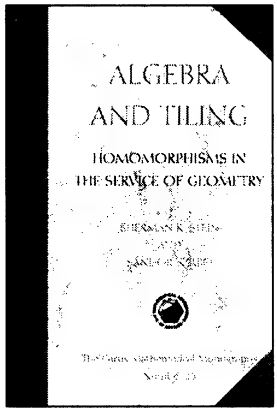
Credit Card No. ______________ Expires ___/___

Signature ______________________

Algebra and Tiling

Homomorphisms in the Service of Geometry

Sherman Stein and Sándor Szabó



Sándor Szabó

Sherman Stein

Algebra and Tiling is perfect for bringing alive an abstract algebra course. Intuitive but difficult problems of geometry are translated into algebraic problems more amenable to solution. Full of nice surprises, the book is a pleasure to read.

—Choice

Often questions about tiling space or a polygon lead to other questions. For instance, tiling by cubes raises questions about finite abelian groups. Tiling by tripods or crosses raises questions about cyclic groups. From tiling a polygon with similar triangles, it is a short step to investigating automorphisms of real or complex fields. Tiling by triangles of equal areas soon involves Sperner's lemma from topology and valuations from algebra.

The first six chapters of *Algebra and Tiling* form a self-contained treatment of these topics, beginning with Minkowski's conjecture about lattice tiling of Euclidean space by unit cubes, and concluding with Laczkowicz's recent work on tiling by similar triangles. The concluding chapter presents a simplified version of Rédei's theorem on finite abelian groups: if such a group is factored as a direct product of subsets, each containing the identity element, and each of prime order, than at least one of them is a subgroup. A remarkable geometric implication of this result is developed in Chapter 2.

Algebra and Tiling is accessible to undergraduate mathematics majors, as most of the tools necessary to read the book are found in standard upper division algebra courses, but teachers, researchers and professional mathematicians will find the book equally appealing. Beginners will find the exercises and the material found in the appendices especially useful. The "Problems" section will appeal to both beginners and experts in the field. The book could serve as the basis of an undergraduate or graduate seminar or a source of applications to enrich an algebra or geometry course.

Contents

224 pp., Hardcover, 1994 ISBN 0-88385-028-1
List: $41.50 MAA Member $33.50
Catalog Code: CAM-25/JV

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ________________

Address ________________

City ________________

State ________ Zip ________

QTY.	CATALOG CODE	PRICE	AMOUNT
____	CAM-25/JV	____	____
____	____	____	____
		TOTAL	____

Payment ☐ Check ☐ VISA ☐ MasterCard

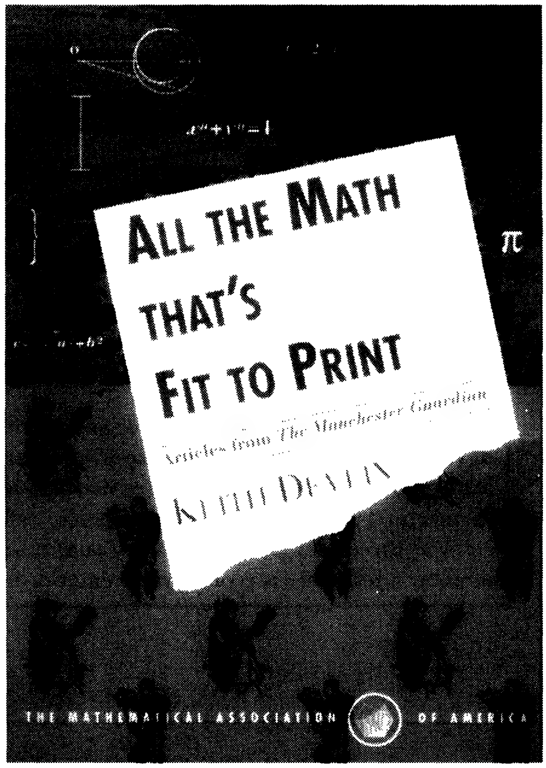
Credit Card No. ________________ Expires ___/___

Signature ________________

All the Math That's Fit to Print

Articles from the Manchester Guardian

Keith Devlin



Mathematics and mathematicians can be the objects of public interest, if there are individuals capable of explaining those items in a form that the intelligent reader can follow. Keith Devlin is such a person and the editors of the British paper, The Manchester Guardian, were intelligent enough to understand that. This book should be an element of every public library.

—Journal of Recreational Mathematics

This new work reveals another side of Devlin's interesting investigations into mathematics and his efforts to share them with laypersons...Anyone interested in mathematics will find something of interest in this book...When possible, the author provides a historical context for the new ideas being explored.

—Choice

Between 1983 and 1989 Keith Devlin, research mathematician, author and educator, wrote a semi-monthly column on mathematics and computing in the English national daily newspaper, The Manchester Guardian. This book is a compilation of many of those articles. It is a witty, entertaining, easy-to-read piece of work.

The mathematical topics range from simple puzzles to deep results including open problems such as Faltings Theorem and the Riemann Conjecture. You will find articles on prime numbers, how to work out claims for traveling expenses, calculating pi, computer simulation, patterns and palindromes, cryptology, and much more.

This book is meant for browsing by anyone who regularly reads a serious newspaper and has some interest in matters scientific or mathematical.

345 pp., Paperbound, 1994
ISBN - 0-88385-515-1
List: $38.50
MAA Member: $28.95
Catalog Code: ATMA/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code: ____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	ATMA/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

SPRINGER FOR MATHEMATICS

PETER HILTON, SUNY at Binghamton, NY; **DEREK HOLTON**, University of Otago, New Zealand and **JEAN PETERSON**, Santa Clara University, CA

MATHEMATICAL REFLECTIONS

In a Room with Many Mirrors

This book shows what mathematics is about, how it is done, and what it is good for. The relaxed and informal presentation conveys the joy of mathematical discovery and insight and makes it clear that mathematics can be an exciting and engrossing activity. The text presents topics, drawn from both pure and applied mathematics, that serve to illustrate the unity of mathematical thought, as well as the diversity of mathematical ideas. Presenting many recent discoveries that lead to interesting open questions, the book can serve as the main text in courses dealing with contemporary mathematical topics or as enrichment for other courses. It can also be read with pleasure on its own by anyone interested in the intellectually intriguing aspects of mathematics.

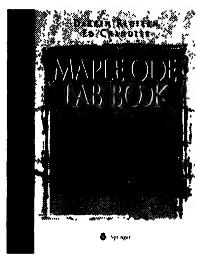
1996/APP. 250 PP., 125 ILLUS.
HARDCOVER/$34.00/ISBN 0-387-94770-1
UNDERGRADUATE TEXTS IN MATHEMATICS

New Second Edition
RICHARD A. HOLMGREN, Allegheny College, Meadville, PA

A FIRST COURSE IN DISCRETE DYNAMICAL SYSTEMS

A discrete dynamical system can be characterized as an iterated function. Given the efficiency with which computers can do iteration, anyone can generate beautiful images whose roots lie in discrete dynamical systems. Images of Mandelbrot and Julia sets abound in publications – the mathematics behind the pictures are beautiful in their own right and are the subject of this text. Mathematica programs that illustrate the dynamics and will aid the student in doing the exercises, are included in the appendix. In this second edition, the covered topics are rearranged to make the text more flexible; it includes the simplification of several proofs; a thorough review and expansion of the exercises; and substantial improvement in the efficiency of the Mathematica programs.

1996/248 PP., 56 ILLUS.
SOFTCOVER/$29.95/ISBN 0-387-94780-9
UNIVERSITEXT



DARREN REDFERN, Practical Approach Corp., Ontario, Canada and **EDGAR CHANDLER**, Paradise Valley Community College

MAPLE ODE LAB BOOK

This book and disk package supplements any introductory text on Ordinary Differential Equations, using Maple V as a computational tool to further understanding. Maple allows one to do many mundane tasks quickly and efficiently, enabling new ways of illustrating and thinking about difficult concepts which will allow for new depths of understanding. Numerical and graphical information and methods are emphasized along with symbolic results in order to take advantage of the versatility of Maple. The manual begins with a well-rounded introduction to Maple so that a background in Maple is not necessary. The worksheets on the disk correspond to the material in each chapter. They provide students with an environment where they can easily experiment with the types of commands found in the text. The manual shows a variety of methods, and the student is encouraged to try and compare different methods of solution.

1996/APP. 150 PP., 50 ILLUS., 3.5" DISK
SOFTCOVER/$29.95 (TENT.)/ISBN 0-387-94733-7

Order Today!

Call: 1-800-SPRINGER
Write: Send payment plus $3.00 postage and handling for the first book and $1.00 for each additional book to: Springer-Verlag New York, Inc., Dept. #S245, PO Box 2485, Secaucus, NJ 07096-2485 (CA, IL, MA, NJ, NY, PA, TX, VA, and VT residents add sales tax, Canadian residents add 7% GST)
Visit: Your local technical bookstore
E-mail: orders@springer-ny.com
Reference: S245 8/96

Springer
http://www.springer-ny.com

COMAP, Inc.
WALTER MEYER, Adelphi University (Coordinating Editor)

PRINCIPLES AND PRACTICE OF MATHEMATICS

This text represents a new entry level course in mathematics for students in programs which will require additional courses in mathematics. Written at approximately the level of introductory calculus, it is an alternative point of entry into the undergraduate mathematics curriculum. By emphasizing the breadth and variety of modern mathematical inquiry and applications, it provides a view of the subject that is not experienced by students in the traditional calculus course. Developed under the direction of COMAP, with NSF support, this text fits into various course arrangements, both for one and two semesters.

ADVISORY BOARD: SAUL GASS, ANDREW GLEASON, JOSEPH MALKEVITCH, DAVID MOORE, PAUL SALLY, LAURIE SNELL, GAIL YOUNG

(Instructors: For an exam copy of the COMAP textbook, please contact Jones & Bartlett Publishers at 1-800-832-0034 or email: springer@jbpub.com)

1996/APP. 816 PP., 400 ILLUS.
HARDCOVER/$57.95 (TENT.)/ISBN 0-387-94612-8
TEXTBOOKS IN MATHEMATICAL SCIENCES

WILLIAM E. FENTON, Bellarmine College, Louisville, KY and **ED DUBINSKY**, Purdue University, West Lafayette, IN

INTRODUCTION TO DISCRETE MATHEMATICS WITH ISETL

Intended for first- or second-year undergraduates, this text applies constructivist principles that promote active participation by the student. Working with the programming language ISETL, whose syntax is close to that of standard mathematical language, the student constructs the concepts in her or his mind as a result of constructing them on the computer in the syntax of ISETL. This approach allows students to discover concepts in a "Socratic" dialog with the computer. The discussion avoids the formal "definition-theorem" approach and promotes active involvement of the reader. An instructor using this text can expect a lively class whose students develop a deep conceptual understanding rather than simply manipulative skills.

1996/APP. 224 PP./HARDCOVER/$39.00
ISBN 0-387-94782-5

THE AMERICAN MATHEMATICAL
MONTHLY

Volume 103, Number 9 November 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT. Periodicals postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership/Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

A Hundred Years of Prime Numbers

Paul T. Bateman and Harold G. Diamond

EARLY WORK ON PRIMES. This year marks the hundredth anniversary of the proof of the Prime Number Theorem (PNT), one of the most celebrated results in mathematics. The theorem is an asymptotic formula for the counting function of primes $\pi(x) := \#\{p \leq x: p \text{ prime}\}$ asserting that

$$\pi(x) \sim x/\log x. \tag{PNT}$$

The twiddle notation is shorthand for the statement $\lim_{x \to \infty} \pi(x)/\{x/\log x\} = 1$. Here we shall survey early work on the distribution of primes, the proof of the PNT, and some later developments.

Since the time of Euclid, the primes, $2, 3, 5, 7, 11, 13, \ldots,$ have been known to be infinite in number. They appear to be distributed quite irregularly, and early attempts to find a closed formula for the nth prime were unsuccessful. By the end of the 18th century many mathematical tables had been computed, and examination of tables of prime numbers led C. F. Gauss and A. M. Legendre to change the question under investigation. Instead of seeking an exact formula for the nth prime, they considered the counting function $\pi(x)$ and asked for approximations to this function, evidently a new kind of question in number theory. Each of the two men conjectured the PNT, though neither did so in the form we have given. In 1808 Legendre published the formula $\pi(x) = x/(\log x + A(x))$, where $A(x)$ tends to a constant as $x \to \infty$. Gauss recorded his conjecture in one of his favorite books of tables around 1792 or 1793 but first disclosed it, in a mathematical letter, over fifty years later. He actually found a better approximation for $\pi(x)$ in terms of the logarithmic integral function, defined for $x > 0$ by

$$\operatorname{li}(x) := \lim_{\epsilon \to 0+} \left\{ \int_0^{1-\epsilon} + \int_{1+\epsilon}^{x} \right\} \frac{1}{\log t}\, dt.$$

It is easy to show that $\operatorname{li}(x) \sim x/\log x$, so either expression can be used in the asymptotic formula for $\pi(x)$. It has been shown that $\operatorname{li}(x)$ is a more accurate estimate of $\pi(x)$ than either $x/\log x$ or Legendre's proposed formula, so today $\operatorname{li}(x)$ is used in PNT error estimates. For more details about Gauss' meditations on the PNT, see **[Gol]**.

The function that we now call the Riemann zeta function, which was to play a decisive role in the proof of the PNT, was introduced by L. Euler in the 18th century. For s real and $s > 1$, define

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$

Using the unique factorization of positive integers, Euler proved that

$$\zeta(s) = \prod_p \{1 + p^{-s} + p^{-2s} + \cdots\} = \prod_p \{1 - p^{-s}\}^{-1},$$

where the product extends over all primes p. Further, he gave another proof of the infinitude of the primes by observing that if the number of primes were finite, then the product for $\zeta(1)$ would converge, while in fact the sum for ζ at $s = 1$ is the harmonic series, which diverges. Euler's proof shows further that the primes are sufficiently numerous that the sum of their reciprocals diverges.

Legendre conjectured and incorrectly believed he had proved that there are an infinite number of primes in each arithmetic progression for which the first term and common difference are relatively prime. This theorem was established by P. L. Dirichlet in 1837 by greatly extending the method of Euler described above. In two papers, Dirichlet introduced characters (periodic completely multiplicative arithmetic functions) to select the elements of an arithmetic progression; he generalized the ζ function by multiplying terms of the series for ζ by characters to make what we today call Dirichlet L functions; he related the value of an L function $L(1, \chi)$ with the class number of quadratic forms of a given discriminant, and from the positivity of the class number he deduced his key lemma that each of the L functions is nonzero at the point $s = 1$. The subject of analytic number theory is generally considered to have begun with Dirichlet.

P. L. Chebyshev

The first person to establish the true order of $\pi(x)$ was P. L. Chebyshev. In the middle of the 19th century he found an ingenious elementary method to estimate

$\pi(x)$ and established the bounds $.921x/\log x < \pi(x) < 1.106x/\log x$ for all sufficiently large values of x. Chebyshev's work was based on use of the arithmetic identity

$$\sum_{d|n} \Lambda(d) = \log n,$$

where von Mangoldt's function Λ is a weighted prime and prime power counting function defined by $\Lambda(d) = \log p$ if $d = p^{\alpha}$ for some prime p and positive integer α and $\Lambda(d) = 0$ otherwise. Chebyshev's formula is the arithmetic equivalent of the zeta function identity $\{-\zeta'(s)/\zeta(s)\} \cdot \zeta(s) = -\zeta'(s)$. Chebyshev showed also that if $\pi(x)/\{x/\log x\}$ had a limit as $x \to \infty$, then its value would be 1. Attempts at improving Chebyshev's methods led to slightly sharper estimates and much more elaborate calculations, but the PNT was not to be established by an elementary method for another hundred years.

A few years after the appearance of Chebyshev's paper, a path to the proof of the PNT was laid out by G. F. B. Riemann **[Edw]**, **[Lan]** in his only published paper on number theory. Riemann's revolutionary idea was to consider ζ as a function of a complex variable and express $\pi(x)$ in terms of a complex integral involving ζ. By formally deforming the integration contour, Riemann achieved an explicit

B. Riemann

formula for $\pi(x)$ as an infinite series whose leading term was $\mathrm{li}(x)$ and that involved the zeros of $\zeta(s)$. However, there was not enough analysis available at that time to rigorously deduce the PNT following Riemann's program. It was not until the end of the 19th century that the missing essential ingredient was supplied: this was the theory of entire functions of finite order, which was developed by J. Hadamard for the purpose of proving the PNT.

Riemann proved that the ζ function has an analytic continuation to **C** with just one singularity, a simple pole with residue 1 at the point $s = 1$ and that ζ satisfies a functional equation connecting its values at complex arguments s and $1 - s$. Incidentally, we owe to Riemann the unusual notation for a complex number $s = \sigma + it$ that has become standard in analytic number theory. Riemann recognized the key role that zeros of the ζ function play in prime number theory. He conjectured several properties of these zeros, all but one of which were proved around the end of the 19th century by Hadamard and H. von Mangoldt. The one conjecture that remains to this day, and is generally considered to be the most famous unsolved problem in mathematics, is the so-called Riemann hypothesis:

$$\text{All nonreal zeros of the } \zeta \text{ function have real part } 1/2. \tag{RH}$$

Riemann evidently perceived the greater difficulty of the RH, for while he stated his other conjectures with no qualification, he prefaced the statement of the RH with the phrase "it is very likely that [es ist sehr wahrscheinlich dass]... ."

Activity in prime number theory increased toward the end of the 19th century. The term "Prime Number Theorem" appears to have originated at this time in the Göttingen dissertation of H. von Schaper "Über die Theorie der Hadamardschen Funktionen und ihre Anwendung auf das Problem der Primzahlen," 1898. There were several false starts before correct proofs of the PNT were given. For example, in 1885 Stieltjes [**Stl**] claimed to have proved the RH. With this result one could establish the PNT with an essentially optimal error term

$$\pi(x) - \mathrm{li}(x) = O(x^{\frac{1}{2}+\epsilon}). \tag{1}$$

Here we have used the notation $f(x) = O(g(x))$, where g is a positive function for all x from some point onward, if $|f(x)|/g(x) < B$ holds for some positive constant B and all sufficiently large positive values of x. The deduction of (1) under the assumption of the RH was later carried out by von Koch. Stieltjes died in 1894 without having either substantiated or retracted his claim of having proved the RH.

FIRST PROOFS OF THE PRIME NUMBER THEOREM. The PNT was established in 1896 by Jacques Hadamard and by Charles-Jean de la Vallée Poussin. It was the first major achievement for each at the start of long and distinguished careers. Hadamard was born at Versailles, France, in 1865. After studies at the Ecole Normale Supérieure he obtained his doctorate in 1892. He spent most of his career in Paris, working principally in complex function theory, partial differential equations, and differential geometry. He died in 1963, within two months of his 98th birthday. De la Vallée Poussin was born in 1866 in Louvain, Belgium, where his father was a professor of mineralogy and geology at the University. After studying at Louvain, he too joined the faculty of the University, at the age of 26, as Professor of Mathematics. His elegant and lucid *Cours d'Analyse* has educated generations of mathematicians in the methods of Borel and Lebesgue. De la Vallée Poussin died in 1962, in his 96th year.

Ch. J. de la Vallée Poussin

The arguments of both Hadamard and de la Vallée Poussin followed the scheme laid out by Riemann. Both papers made essential use of Riemann's functional equation for the zeta function, several other properties of ζ conjectured by Riemann and established by Hadamard, and Hadamard's new theory of entire functions.

Hadamard's paper on the PNT **[Had]** consists of two parts. Here are the opening paragraphs of Part I, "On the distribution of zeros of the zeta function" (in our translation). It is interesting to see how he treats Stieltjes' claim.

> The Riemann zeta function is defined, when the real part of s is greater than 1, by the formula
>
> $$\log \zeta(s) = -\sum_{p} \log(1 - 1/p^s), \tag{2}$$
>
> where p runs over the prime numbers,.... [Translators' remark: Use the principal branch for the logarithms on the right side of (2).] It is holomorphic in the entire plane, except at the point $s = 1$, which is a simple pole. It does not vanish for any value of s with real part greater than 1, since the right-hand side of (2) is finite. But it admits an infinity of complex zeros with real part between 0 and 1. Stieltjes proved, in accordance with Riemann's

expectations, that these zeros are all of the form $\frac{1}{2} + it$ (where t is real); but his proof has never been published, and it has not even been established that the function ζ has no zeros on the line $\Re s = 1$.

It is this last assertion that I propose to prove here.

J. Hadamard

Hadamard's proof that $\zeta \neq 0$ on the line $L\{\sigma = 1\} := \{s \in \mathbf{C}\colon \Re s = 1\}$ used formula (2) for $\log \zeta(s)$, where $s = \sigma + it$ with $\sigma > 1$ and t real, and the representation

$$-\Re \log(1 - p^{-s}) = \Re \sum_{m=1}^{\infty} \frac{1}{mp^{ms}} = \sum_{m=1}^{\infty} \frac{\cos(mt \log p)}{mp^{m\sigma}}.$$

Thus

$$\log|\zeta(s)| = \sum_{p} \sum_{m=1}^{\infty} \frac{\cos(mt \log p)}{mp^{m\sigma}}. \tag{3}$$

In the analysis of ζ, one can ignore the contribution of the higher prime powers, because that part of the series is uniformly bounded for $\sigma > 1$, while the sum over just the primes in (3) is not. Hadamard observed first that, because of the simple

pole of ζ at $s = 1$,

$$\sum_p p^{-\sigma} \sim \log \zeta(\sigma) \sim \log \frac{1}{\sigma - 1} \quad (\sigma \to 1+). \tag{4}$$

He next noted that if $1 + it_0$ were a zero of ζ, necessarily simple, then it would follow that

$$\sum_p p^{-\sigma} \cos(t_0 \log p) \sim -\log \frac{1}{\sigma - 1} \quad (\sigma \to 1+). \tag{5}$$

Comparing (4) and (5), he concluded in succession that (a) $\cos(t_0 \log p) \approx -1$ for most primes p, (b) hence $\cos(2t_0 \log p) \approx +1$ for most primes p, and (c) finally $1 + 2it_0$ would be a pole of ζ, contradicting the fact that ζ has no singularities in $\mathbf{C}$ other than at $s = 1$. We have omitted some details that Hadamard gave to make this argument complete; they can be found also in [**THB**, Ch. 3]. Today it is customary to use a cleaner method, due to F. Mertens, that combines formula (3) with a trigonometric inequality to get an inequality for ζ that expresses Hadamard's idea. For example, the choice $3 + 4\cos\theta + \cos 2\theta \geq 0$ yields

$$\zeta(\sigma)^3 |\zeta(\sigma + it)|^4 |\zeta(\sigma + 2it)| > 1 \quad (\sigma > 1).$$

Part II of Hadamard's 1896 paper, "Arithmetic Consequences," contains his deduction of the PNT. It begins with the following modest words:

> As one can see, we are quite far from having proved the assertion of Riemann-Stieltjes; we have not even been able to exclude the hypothesis of an infinity of zeros of $\zeta(s)$ approaching arbitrarily close to the limiting line $\Re s = 1$. However, the result which we have obtained suffices by itself to prove the principal arithmetic consequences which people have, up to now, sought to deduce from the properties of $\zeta(s)$.

Here are the main ingredients in Hadamard's deduction of the PNT. He firs established the following "smoothed" form of the Mellin inversion formula,

$$\sum_{n<x} a_n \log(x/n) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} \frac{x^s}{s^2} \sum_{n=1}^{\infty} \frac{a_n}{n^s}\, ds,$$

valid for x positive and $\Sigma a_n n^{-s}$ a Dirichlet series that is absolutely convergent fo $\Re s > 1$. The arithmetic function to which he applied the formula was the voı Mangoldt function $\Lambda(n)$ that appeared in Chebyshev's work. The associate(Dirichlet series satisfies the zeta function formula

$$\sum_{n=1}^{\infty} \Lambda(n) n^{-s} = -\zeta'(s)/\zeta(s),$$

which is shown by differentiating formula (2) for $\log \zeta(s)$. Using the Weierstrass Hadamard product representation for $(s-1)\zeta(s)$, the convergence of $\Sigma|\rho|^{-}$ (where ρ runs over the nonreal zeros of ζ), and a contour deformation an(estimation of the above Mellin integral, Hadamard deduced that

$$\sum_{n \leq x} \Lambda(n) \log(x/n) \sim x.$$

From this relation the PNT follows quite easily.

Like Hadamard, de la Vallée Poussin [**VaP**] began his proof by establishing that ζ has no zeros with real part 1 (by a rather more complicated argument than that

of Hadamard). He also used a smoothed form of the Mellin inversion formula, but with an expression $x^s/(s-u)(s-v)$ in place of x^s/s^2. In 1899, de la Vallée Poussin published another article in which he obtained the PNT with an error estimate

$$\pi(x) - \mathrm{li}(x) = O(x \exp\{-c \log^{\alpha} x\}), \tag{6}$$

where $\alpha = 1/2$ and c is some positive constant. In the last paper he made use of Mertens' trigonometric inequality. A quarter of a century went by before de la Vallée Poussin's error bound was improved.

We note that the estimate (6) with a fixed positive value of α is superior to any estimate of the form

$$\pi(x) - \mathrm{li}(x) = O(x/\log^k x)$$

with fixed $k > 0$. The RH implies that (6) holds with $\alpha = 1$, $c = \frac{1}{2} - \epsilon$, as stated in (1).

LATER DEVELOPMENTS. In just over a decade after the proof of the PNT, prime number theory moved from obscurity to mainstream. So little was known on the subject in England at the turn of the century that J. E. Littlewood was assigned the task of proving the RH by E. W. Barnes, his Cambridge research supervisor, and at one point, according to G. H. Hardy, it was believed that the RH had been

E. Landau

proved. The publication of E. Landau's *Handbuch der Lehre von der Verteilung der Primzahlen* [**Lan**] in 1909 quickly changed the status of the subject. Landau's book presented in accessible form nearly everything that was then known about the distribution of primes. Incidentally, the O notation we use was popularized by Landau.

In addition to writing about prime number theory, Landau made significant contributions to the subject, including the simplification of some of the main arguments and extension of the results. For example, he was the first to prove the PNT without making use of the functional equation of ζ. His idea was to combine an analytic continuation of the zeta function a bit to the left of $L\{\sigma = 1\}$, e.g., via

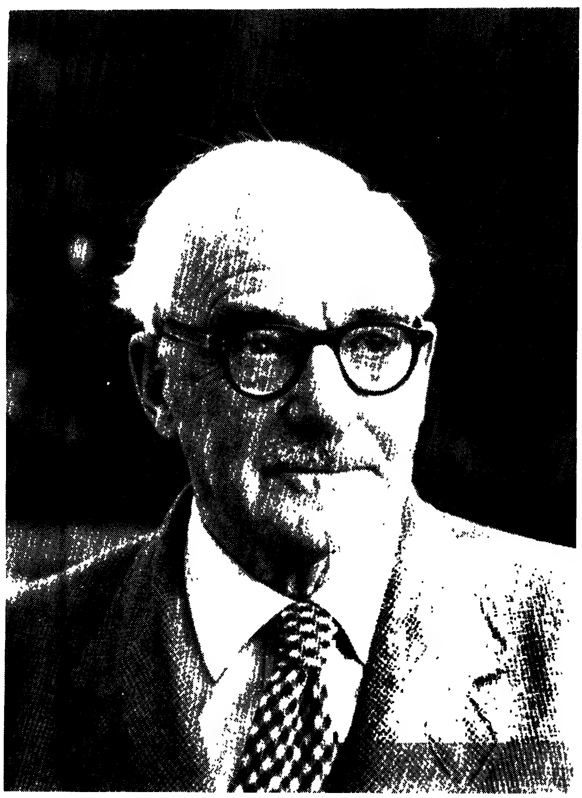
$$\zeta(s) - \frac{s}{s-1} = s\int_1^{\infty} \frac{[x] - x}{x^{s+1}}\,dx, \quad \Re s > 0,$$

with an upper bound for the logarithmic derivative of the zeta function in a suitable zero-free region. With the aid of his new methods, Landau was able to treat some related problems, such as estimating the number of prime ideals of norm at most x in the ring of integers of an arbitrary algebraic number field [**Lan**, Sec. 242]. This result solved part of the eighth problem posed in Hilbert's famous 1900 address to the International Congress of Mathematicians.

It had long been noted, possibly already by Gauss, that

$$\pi(x) - \mathrm{li}(x) < 0 \tag{7}$$

for $x = 2, 3, \ldots$ to whatever point it was checked. In addition to this empirical evidence, theoretical support for the conjecture that (7) holds for all $x \geq 2$ was



J. E. Littlewood

provided by Riemann, who observed that his formula for $\pi(x)$ begins with the terms $\mathrm{li}(x) - \mathrm{li}(\sqrt{x})/2$. However, this conjecture was disproved by Littlewood **[Ing]**, who used almost periodic functions and diophantine ideas to show that in fact the difference changes sign infinitely often. Littlewood's proof did not provide an estimate of where the first change of sign might be, and this question attracted further attention. The suggestion was raised that the question might be undecidable. However, it was proved by S. Skewes that there is a number $x < \exp\exp\exp\exp 7.705$ for which (7) does not hold. Skewes' number, which is among the largest that have occurred in mathematics, has subsequently been replaced by a more modest number with fewer than 400 decimal digits. There is a moral here: vast amounts of empirical evidence together with a "philosophical" explanation for a mathematical phenomenon are not the same as a proof.

What is the relation between the PNT and the nonvanishing of the Riemann zeta function on $L\{\sigma = 1\}$? It is quite easy to see that the PNT implies that ζ has no zeros on the line. Proofs of the PNT were given first by Landau [**Lan**, Sec. 241] and then by Hardy and Littlewood that used, besides the nonvanishing of ζ on $L\{\sigma = 1\}$, only very weak growth conditions for $\zeta(\sigma + it)$ for $\sigma > 1$ and $|t| \to \infty$. The question arose whether the PNT could be proved using just the fact that ζ has no zeros on $L\{\sigma = 1\}$. This was answered affirmatively around 1930 by work of N. Wiener using Fourier analysis. Wiener created an approximate integral formula for $\pi(x)$ involving a compactly supported smoothing function. The following tauberian theorem **[Ch1]** provides one of the most direct proofs now known for the PNT.

Wiener-Ikehara Theorem. *Suppose f is a non-decreasing real-valued function on $[1, \infty)$ such that*

$$\int_1^\infty |f(u)| u^{-\sigma-1}\, du < \infty$$

for each real $\sigma > 1$. Suppose further that

$$\int_1^\infty f(u) u^{-s-1}\, du = \frac{\alpha}{s-1} + g(s), \quad \Re s > 1,$$

where $\alpha \in \mathbf{R}$ and g is the restriction to $\{s\colon \Re s > 1\}$ of a continuous function on the closed half plane $\{s\colon \Re s \geq 1\}$. Then

$$\lim_{u \to \infty} u^{-1} f(u) = \alpha.$$

In 1937 A. Beurling introduced an abstraction of prime number theory in which multiplicative structure was preserved but the additive structure of integers was dropped. A sequence of real numbers $p_1 \leq p_2 \leq p_3 \leq \cdots$, called "generalized primes," was introduced, and the free abelian semigroup generated from them under multiplication was called the associated sequence of "generalized integers." From the assumption that the counting function of the generalized integers satisfies the condition

$$I(x) = Ax + O(x \log^{-\gamma} x), \quad \gamma > 3/2,$$

an analogue of the PNT was established. Moreover, the condition that $\gamma > 3/2$ was shown to be best possible. A form of the Wiener-Ikehara theorem with a weaker hypothesis on the behavior of the function g near $L\{\sigma = 1\}$ can be used in the proof of Beurling's theorem.

N. Wiener

Generalized prime number theory has several applications, and it has raised interesting new problems. For example, Landau's prime ideal theorem is easily deduced from Beurling's result. Also, there are generalized prime models for which the counting function of generalized integers is quite close to that of the usual integers, but for which the analogue of the RH is false. This means that a successful proof of the RH will require more than just the facts that the positive integers are a multiplicative semigroup and that the counting function of positive integers $[x]$ is close to x; presumably, the additive structure of the integers must be taken into account. More on this topic can be found in the authors' survey article **[BaD]**.

Many different proofs have been given for the PNT. A very concise argument that uses only the analyticity and nonvanishing of $(s - 1)\zeta(s)$ on the closed half plane $\{s\colon \Re s \geq 1\}$ was found by D. J. Newman **[New]**. In place of the Wiener-Ikehara theorem or an application of the Mellin inversion integral, Newman's method uses basic complex function theory to estimate the integral

$$\int \frac{x^s}{s}\left(1 + \frac{s^2}{R^2}\right)\sum_{n=1}^{\infty}\frac{a_n}{n^s}\,ds$$

over a finite contour for large values of R. Some other interesting proofs of the

PNT include that of H. Daboussi, which uses elements of sieve theory, and a method of A. Hildebrand based on the large sieve.

De la Vallée Poussin's PNT error term was improved by Littlewood, who used exponential sum methods to find bounds for Dirichlet series. These estimates led to enlarged regions on which the zeta function is guaranteed to be nonzero and consequently to better PNT estimates. The method was developed and improved by the school of I. M. Vinogradov, leading to the bound in which (6) holds with $\alpha = 3/5 - \epsilon$.

The failure of the Chebyshev methods and the success of Riemann's program in proving the PNT led to the opinion, voiced by Hardy and others, that the PNT could be proved only with the use of the Riemann zeta function. This belief was strengthened by Wiener's proof of the equivalence of the PNT and the nonvanishing of ζ on $L\{\sigma = 1\}$. Inspired by work in sieve theory, A. Selberg developed a kind of weighted analogue of Chebyshev's identity. With this formula and an argument of P. Erdös he succeeded in giving an "elementary" proof of the PNT. Subsequently, Selberg and Erdös each discovered an independent proof. Their arguments are considered elementary in the sense that they do not involve the zeta function, complex analysis, or Fourier methods; however, the methods are quite intricate. Subsequently, elementary estimates were sought for the PNT error term, and by use of higher order analogues of Selberg's formula and more elaborate tauberian arguments, error terms of type (6) with $\alpha = 1/6 - \epsilon$ were achieved. For a survey of the use of elementary methods in prime number theory, see **[Dia]**.

We conclude with a summary of what is now known about the truth of the RH. If the RH is false and ζ has even a single nonreal zero off the critical line $\{s \in \mathbf{C}: \Re s = 1/2\}$, there would be consequences for prime number theory, such as in the quality of the PNT error term. The numerical evidence in support of the RH is very great—by comparing the sign changes of a real-valued equivalent of $\zeta(\frac{1}{2} + it)$ with the zeros predicted by use of the argument principle, van de Lune, te Riele, and Winter showed that the first one and a half billion (!) nonreal zeros of zeta lie on the critical line and are simple. In the 1920's, Littlewood showed that almost all the nonreal zeros lie in any given strip of positive width that contains the critical line. Hardy proved that there were infinitely many zeros of zeta on the critical line, and later Selberg showed that a positive proportion of the nonreal zeros were on the line. Near the end of his life, N. Levinson introduced an efficient zero counting method, which B. Conrey has developed to show that more than $2/5$ of the nonreal zeta zeros are simple and lie on the critical line.

SOURCES. The theory of the distribution of prime numbers is a rich and fascinating topic. In this survey we have had to treat fleetingly or omit entirely many interesting topics. Also, it was not feasible to list the sources for all the facts cited. The following books and articles discuss further topics and provide references to original sources.

Landau's *Handbuch* **[Lan]** remains an excellent introduction to prime number theory and is a reference for virtually all early results in the area. The second edition of the *Handbuch*, edited by the first author, contains information on work up to about 1950 on the distribution of primes. The books of Chandrasekharan **[Ch1]**, **[Ch2]**, Ingham **[Ing]**, and Ellison & Mendes-France **[EMF]** provide very readable introductions to the subject. Titchmarsh & Heath-Brown **[THB]** and Ivić **[Ivc]** are standard references on the Riemann zeta function, and Edwards **[Edw]** provides a historical view of this subject. There are detailed and authoritative encyclopedia articles on prime number theory by Hadamard **[BHM]** and by Bohr

and Cramér [**BoC**]. The recent survey article of W. Schwarz [**Sch**] describes the development of prime number theory in the twentieth century, including several topics that we have not treated. Finally, Ribenboim [**Rib**] provides a kind of Guinness record book about primes and includes extensive references.

REFERENCES

[BHM] P. Bachmann, J. Hadamard, and E. Maillet, Propositions transcendantes de la théorie des nombres, *Encyclopédie des sciences mathématiques* I **17** (1910), 215–387.

[BaD] P. T. Bateman and H. G. Diamond, Asymptotic distribution of Beurling's generalized prime numbers, in *Studies in Number Theory*, W. J. LeVeque, ed., Math. Assoc. America, Washington, DC, 1969, 152–210.

[BoC] H. Bohr and H. Cramér, Die neuere Entwicklung des analytischen Zahlentheorie, *Encyklopädie der mathematischen Wissenschaften* II C 8 (1923), 722–849. (Also in Vol. 3 of Harald Bohr's *Collected Mathematical Works*, Copenhagen, 1952).

[Ch1] K. Chandrasekharan, *Introduction to analytic number theory*, Springer-Verlag, 1968.

[Ch2] K. Chandrasekharan, *Arithmetical functions*, Springer-Verlag, 1970.

[Dia] H. G. Diamond, Elementary methods in the study of the distribution of prime numbers, *Bull. Amer. Math. Soc.* (N.S.) **7** (1982), 553–589.

[EMF] W. J. Ellison and M. Mendès-France, *Les nombres premiers*, Hermann, Paris, 1975.

[Edw] H. M. Edwards, *Riemann's zeta function*, Academic Press, 1974.

[Gol] L. J. Goldstein, A history of the prime number theorem, *Amer. Math. Monthly* **80** (1973), 599–615.

[Had] J. Hadamard, Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques, *Bull. de la Soc. math. de France* **24** (1896), 199–220.

[Ing] A. E. Ingham, *The distribution of prime numbers*, Cambridge Univ. Press, Cambridge, 1932.

[Ivc] A. Ivić, *The Riemann zeta-function*, Wiley-Interscience, 1985.

[Lan] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, 2nd ed. (with an appendix by P. T. Bateman), Chelsea Pub. Co., 1953.

[New] D. J. Newman, Simple analytic proof of the prime number theorem, *Amer. Math. Monthly* **87** (1980), 693–696.

[Rib] P. Ribenboim, *The new book of prime number records*, 3rd ed., Springer-Verlag, 1996.

[Sch] W. Schwarz, Some remarks on the history of the prime number theorem from 1896 to 1960, in *Development of Mathematics 1900–1950*, J.-P. Pier, ed., Birkhäuser, 1994.

[Stl] T. J. Stieltjes, Sur une fonction uniforme, *Comptes Rendus Acad. Sci. Paris* **101** (1885), 153–154.

[THB] E. C. Titchmarsh and D. R. Heath-Brown, *The theory of the Riemann zeta-function*, 2nd ed., Oxford Univ. Press, 1986.

[VaP] C.-J. de la Vallée Poussin, Recherches analytiques sur la théorie des nombres premiers, *Ann. de la Soc. scientifique de Bruxelles* **20** (1896), 183–256 & 281–397.

University of Illinois
Department of Mathematics
1409 W. Green St.
Urbana, IL 61801
diamond@math.uiuc.edu / bateman@math.uiuc.edu

Polytope Projection and Projection Polytopes

Thomas Burger, Peter Gritzmann, and Victor Klee

Imagine yourself as the commander of a space ship. Liftoff was a piece of cake, and since then you have been gliding merrily along. But then comes the bad news: A Klingon ship is approaching, and you must prepare for the attack. More bad news: Your batteries are running low! The good news is that your solar cells are working and you are close to a bright star. Thus you can recharge your batteries, but you must certainly do that as quickly as possible. You analyze the situation. Since the solar cells are distributed evenly over the surface of the ship, you decide that you should rotate the ship so that its "face area" is maximized with respect to the light source (assuming that you are still so far from the star that the incoming rays are practically parallel).

A similar but opposite problem arises when you approach a star that emits harmful radiation. You then want to minimize the exposure to the radiation and therefore to minimize the face area in the direction of the star.

In these problems, you are in control of a body in $\mathbb{R}^3$, and you want to turn the body so as to maximize or minimize its "shadow area" with respect to a particular direction of projection (the direction of the incoming rays). In a mathematically equivalent formulation, you may regard the body as being fixed and then look for a direction that maximizes or minimizes the area of the body's projection on a plane orthogonal to the direction.

Projections belong to the basic tools in many areas of mathematics. While the projection on a given subspace can be expressed as a simple matrix operation applied to the original body, it is not so clear how to find projections that are "optimal" with respect to an application that one may have in mind. Problems of this kind occur in a great variety of situations with a similarly great variety of (more or less explicit) criteria for what is a good projection. Examples include the analysis of statistical, astronomical or linguistic data, and also the design and analysis of algorithms for manifold applications. We do not want to elaborate on these applications here; the goal of this paper really is to present some of the (as we hope the reader will agree) beautiful mathematics underlying the special projection problems of maximizing or minimizing the "shadow area" and their higher-dimensional analogues involving orthogonal projections of a body in $\mathbb{R}^n$ onto an $(n-1)$-dimensional subspace. We assume that the body in question is an n-dimensional convex polytope. When $n = 3$, this seems to be a reasonable assumption in the case of the space ship (see Figure 1).

It is not hard to see that when $n = 2$ (so that we are projecting a convex polygon P onto various lines), the maximum projection-length is equal to P's diameter and the minimum projection-length is equal to P's width (the minimum distance between two parallel supporting lines of P) (see Figure 2). Thus the n-dimensional task considered here is one of several ways of extending to $\mathbb{R}^n$ the classical Euclidean task of computing the diameter and the width of a polygon.

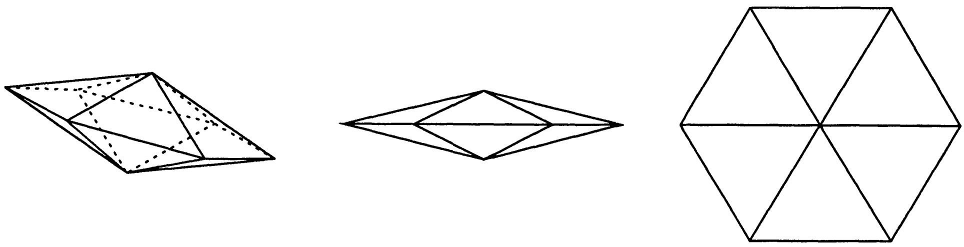


Figure 1. A UFO-shaped polytope seen from different angles.

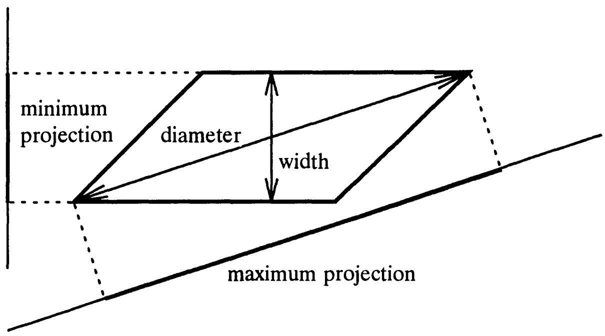


Figure 2. The diameter and width of a polygon are its maximum and minimum projection-lengths.

Polytopes, projections, and volume are all fundamental concepts in *convex geometry*, a mathematical discipline that started with the ground-breaking work of Minkowski (see [21] and [22]). A good advanced introduction to convex geometry from a modern viewpoint is [24], which also contains many historical notes. For accounts focusing on polytopes, [12], [20], and [27] are recommended, and [3] would be a good starting point for undergraduates. Over 60 years old, but still good reading, is [2]. Although we focus here on polytopes, most of the definitions and several of the cited propositions generalize to *convex bodies*, which are defined as nonempty, compact, convex sets. The interested reader may consult [24]. The main questions considered in this paper are the following:

- How can we compute the *shadow volume* (the volume of the orthogonal projection) of a given polytope in a given direction?
- For which direction(s) is the shadow volume a minimum or maximum?

The restriction to orthogonal projections is a natural one, since for maximization we can find arbitrarily large shadow volumes if non-orthogonal parallel projections are admitted. And for minimization, it is easy to see that from a non-orthogonal projection one can obtain an orthogonal projection yielding smaller shadow volume; i.e., minimum projections are orthogonal.

A very useful tool for answering the preceding questions about a polytope P is another polytope, the *projection polytope* Z_P of P. After devoting Section 1 to some basic definitions and the proof of one central theorem, we demonstrate in Section 2 that Z_P provides a straightforward means of computing P's shadow volume for any given direction of projection and is also useful in solving the

minimum problem and the maximum problem. Section 3 discusses these two optimization problems for the case in which P is a simplex, and shows how the diameter and the width of a simplex enter the picture in surprising ways. For a general simplex, minimizing the shadow volume turns out to be solvable in polynomial time, but (even with the aid of the projection polytope) maximizing the shadow volume turns out to be $\mathbb{NP}$-hard. This means that the problem is at least as hard as a number of other difficult problems, such as the Traveling Salesman Problem, for which no efficient algorithm is known or even likely to exist. Such problems are usually called *intractable*. For a detailed account of computational complexity theory see [**7**].

Some of the results and proofs of this paper are well known (see [**2**], [**6**]), some results are provided with proofs that seem to be simpler than those given previously (see [**13**], [**14**], [**16**], [**17**], [**18**], and the survey [**15**]), and still other results appear to be new. However, we would not dare to claim priority for any of these results, because we are convinced that everything was known to Minkowski (with the possible exception of the space ship) but he simply did not get around to writing down all of the relevant elementary observations.

1. POLYTOPES AND THEIR PROJECTIONS. Our general setting is Euclidean n-space $\mathbb{R}^n$ (with $n \geq 2$), equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. The unit sphere $\{u \in \mathbb{R}^n: \|u\| = 1\}$ is denoted by $\mathbb{S}^{n-1}$, and its points are used to represent directions in $\mathbb{R}^n$. The *distance* between two sets $A, B \subset \mathbb{R}^n$ is given by

$$\operatorname{dist}(A, B) = \inf\{\|x - y\|: x \in A, y \in B\}.$$

The dimension of a convex set C is denoted by $\dim C$. The *affine subspaces* of $\mathbb{R}^n$ are simply the translates of the linear subspaces. For an arbitrary subset X of $\mathbb{R}^n$, $\operatorname{lin} X$ and $\operatorname{aff} X$ denote, respectively, the smallest linear subspace and the smallest affine subspace containing X.

An $(n-1)$-dimensional affine subspace is called a *hyperplane*. Each hyperplane divides $\mathbb{R}^n$ into two regions whose closures are called *halfspaces*. We write hyperplanes and halfspaces as

$$H_{x,\alpha} = \{y \in \mathbb{R}^n: \langle x, y \rangle = \alpha\}, \quad \text{respectively, } H_{x,\alpha}^- = \{y \in \mathbb{R}^n: \langle x, y \rangle \leq \alpha\}$$

where $x \in \mathbb{R}^n \setminus \{0\}$, $\alpha \in \mathbb{R}$. Thus, $H_{x,\alpha}$ is the hyperplane orthogonal to x at distance $\alpha/\|x\|$ from the origin. We also write $H_{\pm x,\alpha}^-$ for the "slab" $H_{x,\alpha}^- \cap H_{-x,\alpha}^-$.

A *body* in $\mathbb{R}^n$ is a nonempty compact convex set. We say that a body in $\mathbb{R}^n$ is *proper* if it is n-dimensional. This is equivalent to saying that it has interior points. A *polytope* (as you might have guessed) is a body that is the intersection of a finite number of halfspaces. If P is a proper polytope and

$$P = \bigcap_{i=1}^{m} H_{x_i,\alpha_i}^-$$

for some points $x_i \in \mathbb{R}^m \setminus \{0\}$ and real scalars α_i, and none of the halfspaces can be omitted without enlarging the intersection, then the m sets $P \cap H_{x_i,\alpha_i}$, $i = 1, \dots, m$, are called the *facets* of P. They are $(n-1)$-dimensional polytopes and their union is the boundary of P.

It is well known and rather obvious (although the proof is not so obvious, see, e.g., [**24**, p. 96]) and [**25**, p. 89]) that a body is a polytope if and only if it is the

convex hull of a finite subset of $\mathbb{R}^n$; i.e., a polytope $P \subset \mathbb{R}^n$ can be written as

$$P = \operatorname{conv}\{v_1, \ldots, v_m\} = \left\{ \sum_{i=1}^{m} \lambda_i v_i \colon \lambda_1, \ldots, \lambda_m \in [0, \infty[\wedge \sum_{i=1}^{m} \lambda_i = 1 \right\}$$

for some points $v_1, \ldots, v_m \in \mathbb{R}^n$. If none of the v_i's can be omitted without reducing the convex hull, then the points $v_1, \ldots, v_m$ are the *vertices* of P.

Both the hull representation and the intersection representation of a polytope are important and useful. The n-dimensional cube shows vividly that the sizes of these two representations of a polytope can be drastically different: $2n$ facets versus 2^n vertices (see also [**19**]); see Figure 3 for the cases $n = 2, 3, 4$.

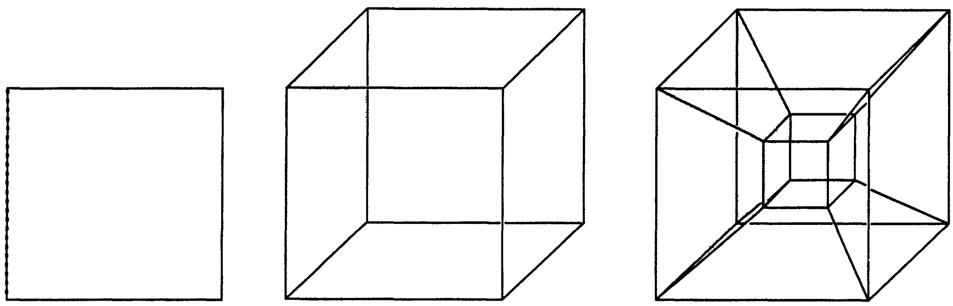


Figure 3. (Projections of) the two-, three-, and four-dimensional cube and some of their facets.

From the description of a polytope as a convex hull, it follows readily that the image of a polytope under any linear transformation is again a polytope. In particular, if P is a proper polytope in $\mathbb{R}^n$ and $u \in \mathbb{S}^{n-1}$, the orthogonal projection of P on the hyperplane $H_{u,0}$ is a polytope of dimension $n - 1$. The $(n - 1)$-volume ($(n - 1)$-dimensional Lebesgue measure) of this projection will be denoted by $\sigma_B(u)$ and called the *shadow volume* of P *in the direction* u.

There is an obvious algorithm for computing the volume of the projection of a polytope on a given subspace: just compute an appropriate representation of the projected polytope in the subspace and then use standard methods (such as triangulation) to compute the volume of the projected polytope. (See [**11**] for a survey of volume computation.) However, a serious drawback of this method is that for each given subspace, a new volume computation is necessary. Wouldn't it be nice to have a simple formula that produces the volume of the projection for each subspace? In fact, there is such a formula, and it leads to a straightforward computation of the shadow volume for any given direction. The formula is closely related to another polytope, the *projection polytope*, that is described in the next section. To prepare for the projection polytope, we consider still another way of describing a proper polytope P in $\mathbb{R}^n$.

Suppose that P is given with facets $F_1, \ldots, F_m$, each facet F_i having $(n - 1)$-volume $\mu_i = V_{n-1}(F_i)$ and lying in some hyperplane H_{u_i, α_i} where $u_i \in \mathbb{S}^{n-1}$ and $\alpha_i \in \mathbb{R}$. To make the orientation of each u_i unique, we assume that u_i is an *outer* (*facet*) *normal vector* of P, meaning that it points outward from P. Formally: for any $x \in F_i$ and any positive λ, we have $x + \lambda u_i \notin P$. We write $\tilde{u}_i = \mu_i u_i$ and call the vectors $\tilde{u}_i$ the *scaled normals* of P. The following theorem leads to the projection polytope.

Theorem 1.1. *Suppose that $F_1, \dots, F_m$ are the facets of a proper polytope $P \subset \mathbb{R}^n$, and $\tilde{u}_1, \dots, \tilde{u}_m$ are the corresponding scaled facet normals. Then $\Sigma_{i=1}^m \tilde{u}_i = 0$. Furthermore, for each $u \in \mathbb{S}^{n-1}$,*

$$\sigma_P(u) = \frac{1}{2} \sum_{i=1}^{m} |\langle u, \tilde{u}_i \rangle|. \tag{1.1}$$

Proof: Let us first consider a facet F_i. For any $u \in \mathbb{S}^{n-1}$, the vector sum $F_i + [0, u]$ is a polytope that may be regarded as a sort of "skewed prism" over the facet F_i. Thinking in terms of F_i, we see that the n-volume of $F_i + [0, u]$ is equal to $V_{n-1}(F_i)$ if $u_i = u$, is equal to 0 if $\langle u, u_i \rangle = 0$, and in general is given by

$$V_n(F_i + [0, u]) = |\langle u, u_i \rangle| V_{n-1}(F_i) = |\langle u, \tilde{u}_i \rangle|.$$

Thinking in terms of the projection, we see that $V_n(F_i + [0, u]) = \sigma_{F_i}(u)$, and thus

$$\sigma_{F_i}(u) = |\langle u, \tilde{u}_i \rangle|.$$

Let us call a facet F_i a *u-facet* if $\langle u, u_i \rangle = 0$. For each such facet, $\sigma_{F_i}(u) = 0$, because the projection of F_i in the direction u is only $(n - 2)$-dimensional. The facets F_i that are not u-facets fall naturally into two classes according to whether $\langle u, u_i \rangle < 0$ or $\langle u, u_i \rangle > 0$; these may be described as the facets that are "visible from the direction u" or "visible from the direction $-u$". The facets in one class form the "u-bottom" of P's boundary with respect to the direction of projection, and those in the other class from the "u-top" (see Figure 4). Since each of the top

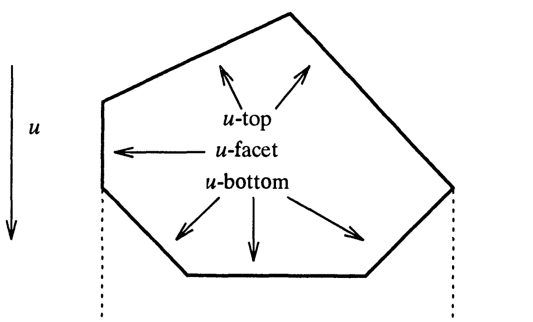


Figure 4. Both the u-bottom and the u-top have the same u-projection as the entire polytope.

and the bottom projects in a one-to-one fashion onto the entire projection of P, we see by summing up over all facets that

$$\sigma_P(u) = \sum_{\langle u, \tilde{u}_i \rangle > 0} \langle u, \tilde{u}_i \rangle \quad \text{and} \quad \sigma_P(u) = - \sum_{\langle u, \tilde{u}_i \rangle < 0} \langle u, \tilde{u}_i \rangle.$$

Subtracting one of these equations from the other, we conclude that $\langle u, \Sigma_{i=1}^m \tilde{u}_i \rangle = 0$ for all $u \in \mathbb{S}^{n-1}$. That yields the first assertion of the theorem. Adding the two equations yields the second assertion.

A famous theorem of Minkowski asserts that whenever $u_1, \dots, u_m$ are m distinct unit vectors whose linear hull is $\mathbb{R}^n$ and $\mu_1, \dots, \mu_m$ are positive numbers such that $\Sigma_{i=1}^m \mu_i u_i = 0$, then there exists an n-polytope P in $\mathbb{R}^n$ that has m facets $F_1, \dots, F_m$ whose scaled facet-normals are the vectors $\tilde{u}_i = \mu_i u_i$. Further, this P is unique up to translation. The algorithmic *reconstruction problem*—producing P from the given data—admits a polynomial-time algorithm for each fixed dimension

n, but the degree of the polynomial grows with n, so that for the case of variable dimension (i.e., when the dimension is "part of the input") the reconstruction problem is #$\mathbb{P}$-hard ([8]). The class #$\mathbb{P}$ contains counting problems, for example the problem of determining the number of Traveling Salesman tours that are shorter than a given bound. This appears to be much more demanding than merely deciding whether the number is positive, and in fact the class of #$\mathbb{P}$-hard problems contains problems that are almost certainly much harder than the $\mathbb{NP}$-complete problems. It is beyond the scope of this paper to give a formal account of complexity theory. However, we do want to emphasize that, although the sense in which algorithmic problems are easy or hard can be made quite precise (see [7]), for the purposes of this paper it suffices to regard as *easy* the problems for which there exists a polynomial-time algorithm, as *hard* the problems that are $\mathbb{NP}$-hard, and as *even harder* or *terribly hard* the problems that are #$\mathbb{P}$-hard.

Determining the scaled normals for a given polytope is also hard in general, for it requires computing the facet volumes (see the survey [11] and the original references given there). However, when the scaled normals $\tilde{u}_i$ are available it's easy to compute P's shadow volume in any given direction u. We merely compute $\frac{1}{2}\sum_{i=1}^{m}|\langle \tilde{u}_i, u\rangle|$ and we're done! But there is more to the story than this, as can be seen in the next two sections.

2. THE PROJECTION POLYTOPE. Another useful way of representing a polytope, and in fact an arbitrary body $B \subset \mathbb{R}^n$, is by means of its *support function* h_B: $\mathbb{R}^n \to \mathbb{R}$. This is defined by setting

$$h_B(x) = \max_{y \in B} \langle x, y \rangle$$

for each $x \in \mathbb{R}^n$. Note that if B is a polytope presented as an intersection of finitely many halfspaces, then for each x the value of $h_B(x)$ can be computed (in polynomial time) by solving a linear program. It can be shown that a convex body is determined by its support function; i.e., two different convex bodies cannot have the same support function.

The intuitive interpretation of the support function h_B is that, for a unit vector $u \in \mathbb{S}^{n-1}$, the hyperplane $H_{u,h_B(u)}$ has nonempty intersection with the boundary of B, and B is contained in the halfspace $H^-_{u,h_B(u)}$; see Figure 5. The hyperplane $H_{u,h_B(u)}$ is called a *supporting hyperplane* of B (in direction u). Thus $h_B(u)$ gives the (signed) distance from the origin of B's supporting hyperplane in the direction

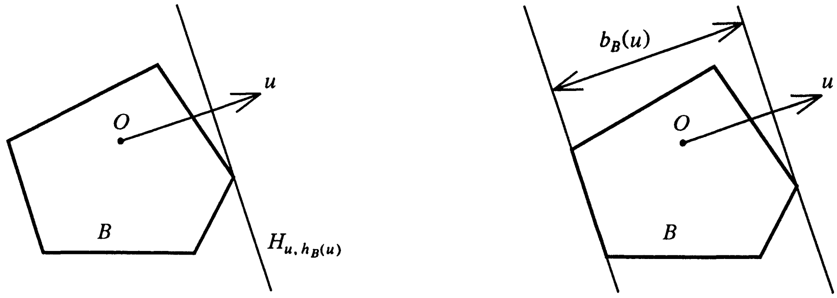


Figure 5. A polytope with a supporting hyperplane resp. two parallel supporting hyperplanes.

u. The closely related *breadth function* $b_B: \mathbb{R}^n \to \mathbb{R}$ is defined by setting

$$b_B(x) = h_B(x) + h_B(-x)$$

for each $x \in \mathbb{R}^n$. For each $u \in \mathbb{S}^{n-1}$ with $\|u\| = 1$, $b_B(u)$ is the distance between the two parallel hyperplanes that are orthogonal to u and "sandwich" B.

Support functions are not always easy to work with, but they do behave nicely with respect to the *Minkowski sum* or *vector sum* $B + C$ of two bodies B and C in $\mathbb{R}^n$, where

$$B + C = \{b + c: b \in B \wedge c \in C\}.$$

It follows immediately from the relevant definitions that

$$h_{B+C} = h_B + h_C.$$

Note also that if B and C are both polytopes, then so is $B + C$. To see this, use the convex hull representation.

Aside from a singleton, the simplest sort of polytope is a *segment*—a set of the form $[x, y] = \operatorname{conv}\{x, y\}$ for distinct points $x, y \in \mathbb{R}^n$. The vector sum of a finite number of segments is a special sort of polytope (called a *zonotope*) that plays an important role here. Note that if $S_1, \dots, S_m$, are segments in $\mathbb{R}^n$, then the sum

$$\sum_{i=1}^{m} S_i = \left\{ \sum_{i=1}^{m} z_i: z_1 \in S_1 \wedge \cdots \wedge z_m \in S_m \right\}$$

is symmetric about the point $\sum_{i=1}^{m} c_i$, where c_i is the center of S_i. For our purposes, it is convenient to assume that the c_i are all at the origin, so we will always consider zonotopes of the form $Z = \sum_{i=1}^{m} [-z_i, z_i]$.

It is easy to see that the support function of a segment $[-z, z]$ is the absolute value of the linear function $f(x) = \langle x, z \rangle$. Thus the following is an immediate consequence of the formula for the support function of a sum of two bodies.

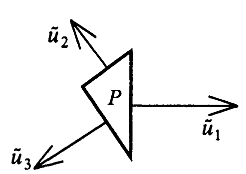
Proposition 2.1. *The support function h_Z of a zonotope $Z = \sum_{i=1}^{m} [-z_i, z_i]$ is given by*

$$h_Z(x) = \sum_{i=1}^{m} |\langle x, z_i \rangle|, \quad x \in \mathbb{R}^n.$$

Zonotopes have many interesting properties, applications, and generalizations for which we have no space here. However, we are at least observing them here in one of their natural habitats, for by comparing Proposition 2.1 and Theorem 1.1 we recognize the right side of (1.1) as the support function of the zonotope

$$Z_P = \frac{1}{2} \sum_{i=1}^{m} [-\tilde{u}_i, \tilde{u}_i].$$

This zonotope Z_P is called the *projection polytope* of P. It is sometimes denoted by ΠP ([24]) but here we have used Z_P to emphasize that it is a zonotope.



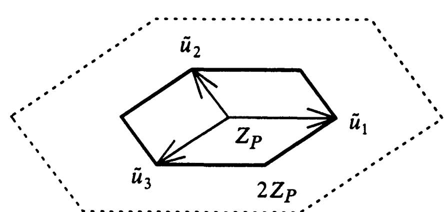


Figure 6. A polytope and its projection polytope.

For a zonotope $Z = \Sigma_{i=1}^{m}[-z_i, z_i]$, each vertex is of the form

$$\sum_{i=1}^{m} \epsilon_i z_i$$

for some choice of $\epsilon_1, \dots, \epsilon_m \in \{-1, 1\}$, and each of its facets is of the form

$$\sum_{i \notin I} \epsilon_i z_i + \sum_{i \in I} [-z_i, z_i]$$

for some choice of the ϵ_i and some index-set $I \subset \{1, \dots, m\}$ such that $\dim \operatorname{lin}\{z_i: i \in I\} = n - 1$. When the zonotope is a *parallelotope* (i.e., when $m = n$ and the vectors $z_1, \dots, z_m$ are linearly independent), all of these choices of ϵ_i and I produce vertices or facets, but otherwise there are choices that do not yield vertices or facets. It is known that the maximum number of k-dimensional faces of an n-dimensional zonotope formed as the sum of m segments is

$$2\binom{m}{k}\sum_{j=0}^{n-1-k}\binom{m-1-k}{j}$$

([4], [26]). In the case of an n-parallelotope, $m = n$ and the formula yields $2n$ facets and 2^n vertices. It follows that for fixed n the maximum number of vertices is $O(m^{n-1})$.

Using the projection polytope we can reformulate the projection problem to obtain the following corollary to Theorem 1.1.

Corollary 2.2. *Let P be a proper polytope in $\mathbb{R}^n$ with scaled normals $\tilde{u}_1, \dots, \tilde{u}_m$, and let I be a proper subset of $\{1, \dots, m\}$ such that $\|\Sigma_{i \in I}\tilde{u}_i\|$ is maximum. Then the shadow volume of P is maximized in the direction of the vector $v = \Sigma_{i \in I}\tilde{u}_i$. That is, the maximum of $\sigma_P(u)$ for $u \in \mathbb{S}^{n-1}$ is attained when $u = (1/\|v\|)v$.*

Proof: Since Z_P is *centered*, i.e., $Z_P = -Z_P$, the maximum of the support function (which, by definition, is the maximum shadow volume) of P equals the circumradius of P. Therefore, it occurs in the direction of a vertex $v_{\max}$ of Z_P most distant from the origin. Any vertex v can be written as $\frac{1}{2}\Sigma_{i=1}^{m}\epsilon_i\tilde{u}_i$ for some $\epsilon_1, \dots, \epsilon_m \in \{-1, 1\}$. Using the fact that $\Sigma_{i=1}^{m}\tilde{u}_i = 0$, we see that

$$v = \frac{1}{2}\sum_{i=1}^{m}\epsilon_i\tilde{u}_i = \frac{1}{2}\sum_{\{i:\, \epsilon_i = 1\}}\tilde{u}_i - \frac{1}{2}\sum_{\{i:\, \epsilon_i = -1\}}\tilde{u}_i = \frac{1}{2}\left(2\sum_{\{i:\, \epsilon_i = 1\}}\tilde{u}_i\right).$$

The stated conclusion follows from this.

This result yields a finite algorithm for the maximum problem: just compute $\Sigma_{i \in I}\tilde{u}_i$ for every proper subset I of $\{1, \dots, m\}$ and take the maximum value obtained. That requires the evaluation of $\Sigma_{i \in I}\tilde{u}_i$ for $2^m - 2$ different choices of the index-set I. By taking account of the symmetry of the problem, this number can be divided by two. In [18], an improved algorithm is given that also uses the above corollary (it is proved there, too). That algorithm considers only subsets I for which $\Sigma_{i \in I}\tilde{u}_i$ is a vertex of Z_P, thus reducing the number of arithmetic operations of the algorithm to $O(m^{n-1})$. In particular, the algorithm has polynomial time-complexity for any fixed dimension n.

We have seen that, since Z_P is centered at the origin, to solve the maximum problem we can alternatively search for a point $z_{\max}$ of Z_P most distant from the origin. Similarly, for the minimum problem, we can also look for a point $z_{\min}$ in the boundary of Z_P closest to the origin. While each $z_{\max}$ is a vertex of Z_P, each

$z_{\min}$ must lie in some facet of Z_P to which it is orthogonal. In conjunction with our earlier description of vertices and facets of zonotopes, this leads to a short proof of a theorem of **[16]** concerning directions of projections that maximize or minimize the shadow volume.

Theorem 2.3. *Let P be a proper polytope in $\mathbb{R}^n$. If the shadow volume of P has a maximum [minimum] in direction $u \in \mathbb{S}^{n-1}$, then no facets [at least $n-1$ facets with linearly independent outer normal vectors] are u-facets.*

Proof: Let $Z_P = \frac{1}{2}\sum_{i=1}^m[-\tilde{u}_i, \tilde{u}_i]$ with $\tilde{u}_1, \ldots, \tilde{u}_m \in \mathbb{R}^n \setminus \{0\}$. Let $z_{\max} = \frac{1}{2}\sum_{i=1}^m \epsilon_i \tilde{u}_i$, where $\epsilon_1, \ldots, \epsilon_m \in \{-1, 1\}$, be a vectex of Z_P having maximum norm. For $j = 1, \ldots, m$, let $z_j = z_{\max} - \epsilon_j \tilde{u}_j$. Of course, $z_j \in Z_P$ and thus

$$\|z_{\max}\|^2 - 2\epsilon_j\langle z_{\max}, \tilde{u}_j\rangle + \|\tilde{u}_j\|^2 = \|z_j\|^2 \le \|z_{\max}\|^2$$

since $\|z_{\max}\|$ is maximum. This means $0 < \|\tilde{u}_j\|^2 \le 2\epsilon_j\langle z_{\max}, \tilde{u}_j\rangle$; i.e., $\langle z_{\max}, \tilde{u}_j\rangle \neq 0$, and the statement about the maximum is proved.

The point $z_{\min}$ is orthogonal to a facet of Z_P. Since Z_P is a zonotope, any facet of Z_P is of the form $\sum_{i \notin I}\epsilon_i\tilde{u}_i + \sum_{i \in I}[-\tilde{u}_i, \tilde{u}_i]$ for some $\epsilon_1, \ldots, \epsilon_m \in \{-1, 1\}$ and $I \subset \{1, \ldots, m\}$, where $\dim \operatorname{lin}\{\tilde{u}_i : i \in I\} = n - 1$. Thus $z_{\min}$ is orthogonal to the (at least $n-1$) vectors $\tilde{u}_i$, $i \in I$.

The preceding characterization of the minimum yields an algorithm that finds the minimum direction, since only $\binom{m}{n-1}$ directions must be considered.

Although the preceding algorithms run in polynomial time when the dimension n is fixed and the scaled facets-normals $\tilde{u}_i$ are given, the scaling factors are the facet-volumes, which in general are difficult to find; see **[11]**. The rest of this section is devoted to parallelotopes, for which the facet volumes are easily computed and hence the shadow volume in any particular direction can be computed in polynomial time. The following theorem gives the projection polytope of a parallelotope. It clearly suffices to consider parallelotopes centered at the origin.

Theorem 2.4. *Let $w_1, \ldots, w_n \in \mathbb{R}^n$ be linearly independent. Then for the parallelotope $P = \bigcap_{i=1}^n H^-_{\pm w_i, 1}$, the projection polytope is*

$$Z_P = \frac{1}{2}V_n(P)\sum_{i=1}^n [-1, 1]w_i.$$

Proof: Let $\pm F_i$ be the facets of P lying in $H_{\pm w_i, 1}$, $\mu_i = V_{n-1}(F_i)$, and $u_i = w_i/\|w_i\|$, $i = 1, \ldots, n$. Then $V_n(P) = 2\mu_i/\|w_i\|$, and therefore $\mu_i u_i = \frac{1}{2}V_n(P)w_i$. By Theorem 1.1 and the definition of the projection polytope, this proves the assertion.

This theorem shows that the projection polytope of a parallelotope is itself a parallelotope. That is good for the minimum problem, because, for a parallelotope centered at the origin and given as the intersection of halfspaces, it is easy to find a point q in the boundary that is closest to the origin: in each hyperplane H containing a facet, find the point q_H of minimum norm, then let q be a q_H that is closest to the origin. We could also use Theorem 2.3 to show that the minimizing direction is orthogonal to $n-1$ of the outer normals. Hence there are only n directions to consider. However, we also see that the maximum problem amounts to finding, for a parallelotope Z_p that is centered at the origin, a vertex that is at maximum distance from the origin. This problem is known to be $\mathbb{NP}$-hard (see **[1]**). We refer to **[5]** for details and more $\mathbb{NP}$-hardness results.

3. SIMPLICES. Simplices are arguably the simplest proper polytopes, for they have the smallest possible number of vertices and facets. In this section we derive the projection polytope of the simplex and a simple relation between the simplex and its projection polytope. This also yields surprisingly simple relations between the maximum and minimum projections and the width and the diameter of a simplex. The maximum problem proves again to be very hard even for this simple polytope.

Let $v_0, \dots, v_n$ be the vertices of a proper simplex S in $\mathbb{R}^n$ with facets $F_0, \dots, F_n$ numbered so that $v_i \notin F_i$. Let $\mu_i = V_{n-1}(F_i)$, let u_i denote the unit outer normal of F_i, and let $\tilde{u}_i = \mu_i u_i$ denote the scaled normals of S, $i = 0, \dots, n$. Note that by our use of indexing, $F_k = \operatorname{conv}\{v_j\colon\ j = 0, \dots, n \wedge j \neq k\}$ for $k = 0, \dots, n$. This means that for all $i, j, k = 0, \dots, n$ with $k \notin \{i, j\}$ the vector $v_i - v_j$ is parallel to $\operatorname{aff}(F_k)$, whence $\langle v_i - v_j, \tilde{u}_k \rangle = 0$. If, on the other hand, $k \in \{i, j\}$, then the set $C = \operatorname{conv}\{F_k, v_i - v_j + F_k\}$ is a (skew) cylinder with base F_k having the same height as S, hence

$$nV_n(S) = V_n(C) = \left|\langle v_i - v_j, \tilde{u}_k \rangle\right|.$$

Summarizing these simple observations, we obtain the following lemma.

Lemma 3.1. *Suppose that S is a proper simplex in $\mathbb{R}^n$ with vertices $v_0, \dots, v_n$ and scaled outer facet-normals $\tilde{u}_i$. Then for $i, j, k \in \{0, \dots, n\}$ it is true that*

$$\left|\langle v_i - v_j, \tilde{u}_k \rangle\right| = \begin{cases} 0 & \textit{if } i = j \textit{ or } k \notin \{i, j\}, \\ nV_n(S) & \textit{otherwise}. \end{cases}$$

With the aid of *polarity* (another standard tool from convex geometry), this lemma yields another useful representation of the projection polytope of a simplex. The *polar* P^* of a polytope $P \subset \mathbb{R}^n$ containing the origin in its interior is

$$P^* = \{x \in \mathbb{R}^n\colon \langle x, y \rangle \leq 1 \text{ for all } y \in P\}.$$

It is not difficult to show that for a polytope P given as the intersection of half-spaces, say $P = \bigcap_{i=1}^m H^-_{x_i, 1}$ for some $x_1, \dots, x_m \in \mathbb{R}^n \setminus \{0\}$, $P^* = \operatorname{conv}\{x_1, \dots, x_m\}$. Furthermore, $(P^*)^* = P$. In particular, the polar of a polytope is again a polytope, and the facets and vertices of a polytope correspond respectively to the vertices and facets of its polar (see [24]).

The *difference polytope* D_P of a polytope P is defined by

$$D_P = P - P = \{x - y\colon x, y \in P\};$$

it is sometimes denoted by DP [23].

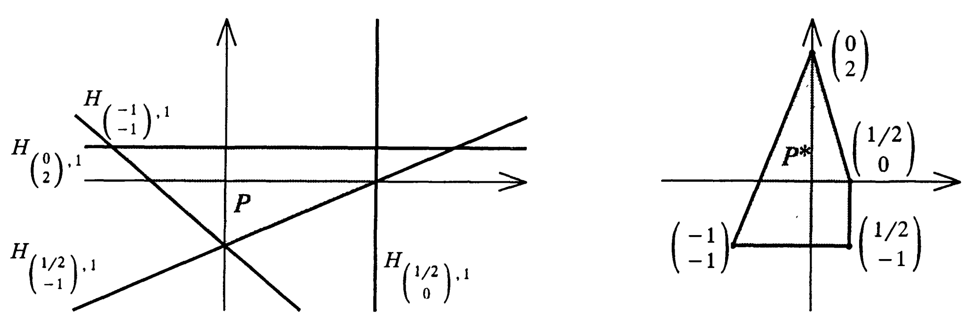


Figure 7. A polytope and its polar.

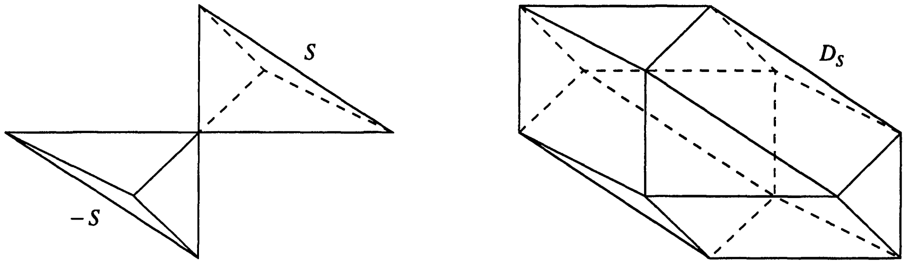


Figure 8. A simplex and its difference polytope.

Theorem 3.2. *The projection polytope Z_S and the difference polytope D_S of a proper simplex $S \subset \mathbb{R}^n$ are related by polarity*:

$$Z_S = nV_n(S)(D_S)^*.$$

Proof: Since $D_S = \operatorname{conv}\{v_i - v_j\colon i, j = 0, \dots, n\}$ we have

$$(D_S)^* = \{x \in \mathbb{R}^n\colon \langle v_i - v_j, x\rangle \leq 1 \text{ for all } i, j = 0, \dots, n\}.$$

The facets of $Z_S = \frac{1}{2}\sum_{i=0}^{n}[-1,1]\tilde{u}_i$ can be obtained as follows. Choose two different scaled normals, say, $\tilde{u}_j$ and $\tilde{u}_k$, and let $I_{j,k} = \{0, \dots, n\} \setminus \{j, k\}$. Then

$$F = \frac{1}{2}(\tilde{u}_j + \tilde{u}_k) + \frac{1}{2}\sum_{i \in I_{j,k}}[-1,1]\tilde{u}_i$$

and $-F$ are facets of Z_S, all such sets are different, and all facets are obtained in this way. By Lemma 3.1, the vector $v_j - v_k$ is normal to F and, furthermore,

$$\left|\left\langle \tfrac{1}{2}(\tilde{u}_j + \tilde{u}_k), v_j - v_k\right\rangle\right| = nV_n(S).$$

Therefore,

$$\operatorname{aff} F \cup \operatorname{aff}(-F) = \{x \in \mathbb{R}^n\colon |\langle v_j - v_k, x\rangle| = nV_n(S)\}.$$

This shows that

$$Z_S = \{x \in \mathbb{R}^n\colon \langle v_i - v_j, x\rangle \leq nV_n(S) \text{ for all } i, j = 0, \dots, n\} = nV_n(S)(D_S)^*$$

which completes the proof.

The preceding polarity result is deduced in **[17]** by means of a different approach. Some other interesting results can be found in **[23]**.

The result of Theorem 3.2 seems rather abstract at first sight, so let us give an additional interpretation involving the *u-length* $l_P(u)$, $u \in \mathbb{S}^{n-1}$, of a polytope P, which is defined as the length of the longest segment in P parallel to u. From Theorem 3.2, we then obtain the following relation between projections and longest segments of simplices in a given direction; see **[17]**, and **[13]**, **[14]** for a proof that this actually characterizes simplices.

Theorem 3.3. *For any simplex $S \subset \mathbb{R}^n$ and direction $u \in \mathbb{S}^{n-1}$, it is true that*

$$l_S(u)\sigma_S(u) = nV_n(S).$$

Proof: The result is obvious if $V_n(S) = 0$. Otherwise, S is proper and $l_S(u) \neq 0$ for all $u \in \mathbb{S}^{n-1}$. By definition,

$$Z_S = \{x \in \mathbb{R}^n\colon \langle x, u\rangle \leq \sigma_S(u) \text{ for all } u \in \mathbb{S}^{n-1}\}.$$

On the other hand, by Theorem 3.2 and the definition of the polar polytope,

$$
\begin{aligned}
Z_S &= \left\{x \in \mathbb{R}^n: \left\langle \frac{x}{nV_n(S)}, y \right\rangle \le 1 \text{ for all } y \in D_S \right\} \\
&= \left\{x \in \mathbb{R}^n: \left\langle \frac{x}{nV_n(S)}, l_S(u)u \right\rangle \le 1 \text{ for all } u \in \mathbb{S}^{n-1} \right\} \\
&= \left\{x \in \mathbb{R}^n: \langle x, u \rangle \le \frac{nV_n(S)}{l_S(u)} \text{ for all } u \in \mathbb{S}^{n-1} \right\}.
\end{aligned}
$$

Together, this yields $\sigma_S(u) = nV_n(S)/l_S(u)$ for all $u \in \mathbb{S}^{n-1}$, which is the asserted relation.

In view of the close relations

$$\max_{u \in \mathbb{S}^{n-1}} l_P(u) = \max_{u \in \mathbb{S}^{n-1}} b_P(u) \quad \text{and} \quad \min_{u \in \mathbb{S}^{n-1}} l_P(u) = \min_{u \in \mathbb{S}^{n-1}} b_P(u)$$

of l_P and b_P (see **[10]**) it is natural to wonder whether there are relations similar to those given in Theorem 3.3 for the breadth. The following result can be found in [**6**, pp. 112–13].

Theorem 3.4. *For $\emptyset \neq I \subsetneqq \{0, \dots, n\}$ and a proper simplex $S \subset \mathbb{R}^n$, let $\tilde{u}_I = \sum_{i \in I} \tilde{u}_i$, where $\tilde{u}_0, \dots, \tilde{u}_n$ denote the scaled normals of S. Then $\tilde{u}_I \neq 0$ and, with $u_I = \tilde{u}_I/\|\tilde{u}_I\|$,*

$$b_S(\tilde{u}_I) = b_S(u_I)\|\tilde{u}_I\| = nV_n(S).$$

Proof: If $\tilde{u}_I$ were zero, then n of the scaled normals would be linearly dependent and thus lie in a hyperplane. Since $\sum_{i=0}^n \tilde{u}_i = 0$, the remaining scaled normal would also lie in this hyperplane, so S would be less than full-dimensional, contrary to our assumption.

With the aid of $\sum_{i=0}^n \tilde{u}_i = 0$ and Lemma 3.1 we see that $\tilde{u}_I$ is a normal of the $(n-1)$-dimensional subspace that leads to the parallel pair of hyperplanes H_1, H_2 with $\operatorname{aff}\{v_i: i \in I\} \subset H_1$ and $\operatorname{aff}\{v_i: i \notin I\} \subset H_2$. Without loss of generality we can assume that $v_1 \in H_1$ and $v_2 \in H_2$. Then, using Lemma 3.1 again,

$$b_S(u_I) = \operatorname{dist}(H_1, H_2) = \frac{|\langle v_1 - v_2, \tilde{u}_I \rangle|}{\|\tilde{u}_I\|} = \frac{|\langle v_1 - v_2, \tilde{u}_1 \rangle|}{\|\tilde{u}_I\|} = \frac{nV_n(S)}{\|\tilde{u}_I\|}$$

as asserted.

Theorem 3.4 relates the breadth $b_S(u_I)$ to the projection of S on a hyperplane that is perpendicular to u_I. Clearly, this result is identical with Theorem 3.3 in the cases of minimum and maximum breadth, respectively, minimum and maximum length. Let us write

$$
\begin{aligned}
l_{\min}(P) &= \min_{u \in \mathbb{S}^{n-1}} l_P(u) = \min_{u \in \mathbb{S}^{n-1}} b_P(u), \\
l_{\max}(P) &= \max_{u \in \mathbb{S}^{n-1}} l_P(u) = \max_{u \in \mathbb{S}^{n-1}} b_P(u).
\end{aligned}
$$

Then, as a corollary to both Theorems 3.3 and 3.4, we obtain the following equalities.

Corollary 3.5. *For a simplex $S \subset \mathbb{R}^n$,*

$$l_{\min}(S) \max_{u \in \mathbb{S}^{n-1}} \sigma_S(u) = nV_n(S) \quad \text{and} \quad l_{\max}(S) \min_{u \in \mathbb{S}^{n-1}} \sigma_S(u) = nV_n(S).$$

This result was used in [**9**] to show that the problem of computing the width of a simplex and the problem of finding a maximum vertex of a parallelotope are polynomially equivalent.

Since, for a simplex S, $l_{\max}(S)$ is just the maximum of the distances between the vertices, this quantity is easy to compute. That yields an algorithm for finding the minimum projection of a simplex. The algorithm performs at most $O(n^2)$ steps, and thus runs in polynomial time even when the dimension is not fixed. For maximization, the situation is much more difficult. In fact, it is possible to show that the problem of finding a maximum projection for simplices is $\mathbb{NP}$-hard, [**5**].

ACKNOWLEDGMENTS. The work of the second and third authors was in part made possible by the Deutsche Forschungsgemeinschaft, by a joint Max-Planck Research Award, and by a NATO Travel Award. The third author's work was also supported in part by the National Science Foundation.

REFERENCES

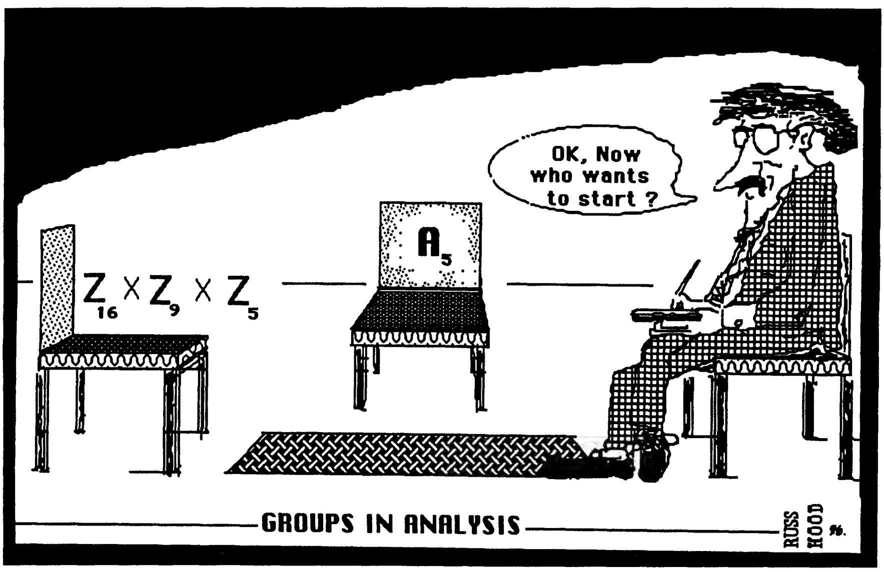
1. H. L. Bodlaender, P. Gritzmann, V. Klee, and J. van Leeuwen, *Computational complexity of norm-maximization*, Combinatorica **10** (1990), 203–225.
2. T. Bonnesen and W. Fenchel, *Theorie der Konvexen Körper*, Springer, Berlin, 1934, Reprint: Chelsea Publ. Co., New York, 1948, Translation: BCS Associates, Moscow, Idaho, 1987.
3. A. Brøndsted, *An Introduction to Convex Polytopes*, Springer, New York, 1983.
4. R. C. Buck, *Partition of space*, Amer. Math. Monthly **50** (1943), 541–544.
5. T. Burger and P. Gritzmann, *Finding optimal shadows of polytopes*, in preparation, 1996.
6. H. G. Eggleston, *Convexity*, Cambridge Univ. Press, 1958.
7. M. R. Garey and D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.
8. P. Gritzmann and A. Hufnagel, *A polynomial time algorithm for Minkowski reconstruction*, Proc. 11th Ann. Symp. Comp. Geom., Association Comput. Mach., 1995, pp. 1–9.
9. P. Gritzmann and V. Klee, *On the 0-1 maximization of positive definite quadratic forms*, Operations Research Proceedings (D. Pressmar et al., eds.), Deutsche Gesellschaft für Operations Research, Springer, Berlin, 1989, pp. 222–227.
10. ______, *Inner and outer j-radii of convex bodies in finite-dimensional normed spaces*, Discrete Comput. Geom. **7** (1992), 255–280.
11. ———, *Some basic problems in computational convexity: II. Volume and mixed volumes*, In: Polytopes: Abstract, Convex and Computational (T. Bisztriczky, P. McMullen, R. Schneider, and A. Ivic Weiss, eds.), Kluwer, Boston, 1994, pp. 373–466.
12. B. Grünbaum, *Convex Polytopes*, Wiley-Interscience, London, 1967.
13. H. Martini, *Some characterizing properties of the simplex*, Geom. Dedicata **29** (1989), 1–6.
14. ______, *A new view on some characterizations of simplices*, Arch. Math. **55** (1990), 389–393.
15. ______, *Cross-sectional measures*, Coll. Math. Soc. J. Bolyai, North-Holland, Amsterdam, 1994, pp. 269–310.
16. H. Martini and B. Weissbach, *Zur besten Beleuchtung konvexer Polyeder*, Beitr., Algebra Geom. **17** (1984), 151–168.
17. ———, *On quermasses of simplices*, Studia Sci. Math. Hung. **27** (1992), 213–222.
18. M. McKenna and R. Seidel, *Finding the optimal shadow of a convex polytope*, Proc. 1st Symp. Comp. Geom., Association Comput. Mach., 1985, pp. 24–28.
19. P. McMullen, *The maximum number of faces of a convex polytope*, Mathematika **17** (1970), 179–184.
20. P. McMullen and G. C. Shephard, *Convex Polytopes and the Upper Bound Conjecture*, Cambridge Univ. Press, 1971.
21. H. Minkowski, *Allgemeine Lehrsätze über die konvexen Polyeder*, Nachr. Ges. Wiss. Göttingen (1897), 198–219.
22. ______, *Volumen und Oberfläche*, Math. Ann. **57** (1903), 447–495.
23. C. A. Rogers and G. C. Shephard, *The difference body of a convex body*, Arch. Math. **8** (1957), 220–233.

24. R. Schneider, *Convex Bodies*: *The Brunn-Minkowski Theory*, Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge Univ. Press, 1993.
25. A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Interscience, New York, 1986.
26. T. Zaslavsky, *Facing up to Arrangements*: *Face-Count Formulas for Partitions of Space by Hyperplanes*, American Mathematical Society Memoirs, vol. 154, Amer. Math. Soc., 1975.
27. G. M. Ziegler, *Lectures on Polytopes*, Graduate Texts in Mathematics, vol. 152, Springer-Verlag, New York–Berlin–Heidelberg, 1995.

Burger:
Universität Trier
Fb IV, Mathematik
D-54286 Trier
Germany
burger@dm7.uni-trier.de

Gritzmann:
Universität Trier
Fb IV, Mathematik
D-54286 Trier
Germany
gritzman@dm1.uni-trier.de

Klee:
University of Washington
Department of Mathematics
Box 354350
Seattle, WA 98195-4350
klee@math.washington.edu



Contributed by Russell H. Hood, California State University, Sacramento.

The Mercedes Knot Problem

Aleksandar Jurišić

1. INTRODUCTION. Have you ever coiled a long extension cord in order to store it (e.g., after you have finished vacuum-cleaning or mowing the lawn)? After you unwind it, you may notice that it is twisted many times (see Figure 1). There is a mathematical problem that is closely related to this phenomenon. It can challenge even a non-mathematician. Imagine a ball in the middle of a room, connected with three elastic bands b, m, and w to the walls. The elastic bands, whose ends are fixed, can be moved in any way but cannot be torn. The question is whether it is possible to move them in such a way that the elastic band m is wrapped several times around the elastic band w without moving the ball (see Figure 2). This is a geometrical problem, somewhere between topology and knot theory, however our solution does not really use any topology and is based on simple combinatorics.

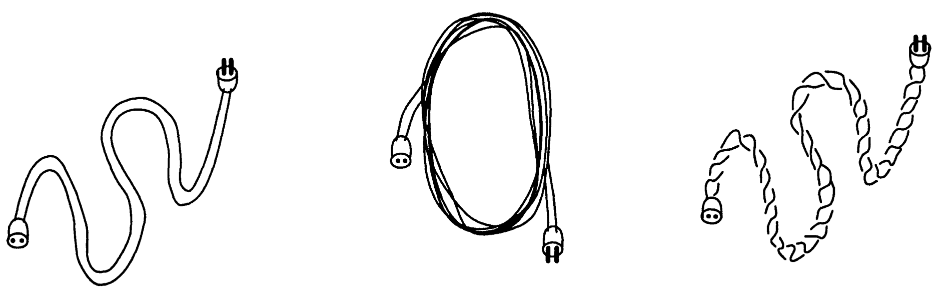


Figure 1. Why do the extension cords usually get twisted?

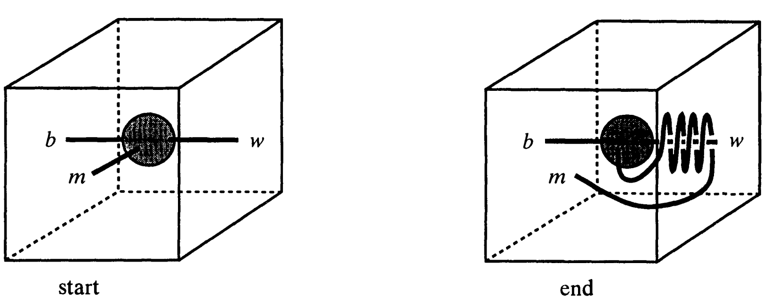


Figure 2

We first give a partial solution, then we translate the problem into mathematical language and give some historical background. In Section 4 we try to understand the gist of the problem by studying a few simpler situations from the knot theory point of view. After that we are able to give a complete solution. Along the way

some interesting connections arise and at the end we mention some applications. Problem solvers will find a few interesting puzzles. We keep in mind Hilbert's advice: "*The art of doing mathematics consists in finding that special case which contains all the germs of generality.*"

2. PARTIAL SOLUTION. In a way this problem is mysterious, since from the very beginning our intuition may lead us in a wrong direction. Let us start with a few naive approaches. If we allowed the ball to rotate around its center, then the answer would trivially be 'yes': we could wrap the elastic band m around w as many times as we wanted (just by rotating the ball around the line determined by the ends of w), and the problem would not be interesting. We could reach a similar conclusion if we assumed for a moment that only the bands w and m were in the room. (Why?) But we cannot rotate the ball and there are three bands in the room, so maybe the answer is 'no'.

On the other hand, if we believe that the answer is 'yes', we may assume that this can be obtained even when the elastic bands b and w are rigid (do not move at all). This belief might be based on the fact that b and w are at the same place in both positions of Figure 2, or that the rotation of the ball would not have moved b and w (but it would twist them). So if the bands b and w are rigid, the space where m can be moved "looks like" a filled torus, which is what mathematicians call a donut (in topological language we would say "is homeomorphic to" $S^1 \times B^2$). But then there is no way to move the band m in a filled torus without moving its ends in order to change the number of times the band m is wrapped around the hole of the filled torus. A topologist would say that the starting and the ending position of m correspond respectively to the identity and a non-identity element of the fundamental group of the filled torus; see Neuwirth [**Neu**]. We conclude that either the movement of the elastic bands b and w is crucial, or the answer is 'no'.

Surprisingly, the answer (yes or no) depends on how many times we want to wrap m around w. Let us first start exploring the question of existence. By building a real model we find out unexpectedly quickly how to wrap the band m twice around the band w; see Figure 3(a), which also explains the title of this paper. As usual in knot theory we prefer to draw two-dimensional figures. They are basically projections (Figure 3(b) shows what happens at crossings) that contain all the necessary information to build the three dimensional objects they describe. Furthermore, the situation remains unchanged if we consider a two-dimensional sphere instead of a cube. An enthusiastic reader is invited to play with her/his own model before verifying our solution in Figure 4. We promise to the reader a better understanding of our solution by the end of this paper.

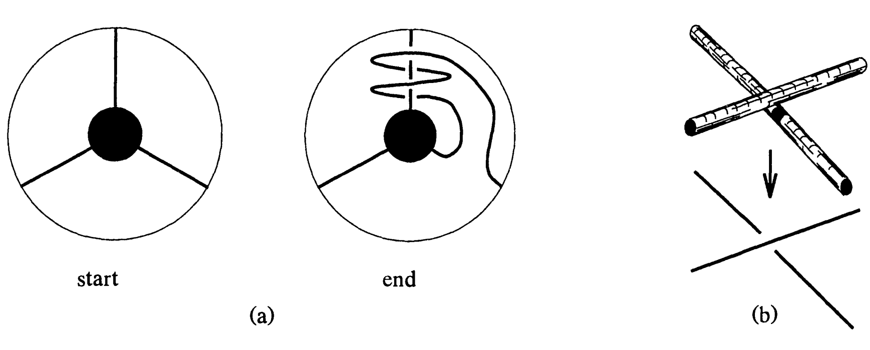


Figure 3

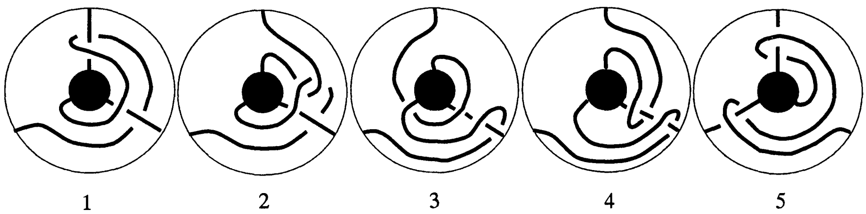


Figure 4. These are the intermediate steps between the above 'start' and 'end' position.

It is now obvious that, by moving the coils to the larger sphere and then repeating the procedure from Figure 4, we can wrap the band m around the band w any even number of times. The problem is much harder if we want to wrap the band m around the band w an odd number of times.

3. MODELING AND HISTORY OF THE PROBLEM. Let us translate the problem into mathematical language. As we are describing a geometrical problem, we will need some terminology from topology, however our solution requires very little topology. So let us follow topologists who prefer to speak about a motion of a space, called an *ambient isotopy*, instead of a movement of objects in it. To understand this better, imagine that the room is filled with honey (or any sticky substance). Then a movement of elastic bands in the room would cause honey to move. On the other hand, a motion of honey could move the elastic bands. So we usually ask for the existence of an ambient isotopy that moves objects from one position to another.

Let S_1 and S_2 be two-dimensional spheres with centers at the origin of $\mathbb{R}^3$ and radii one and two, respectively. Let H be the space between S_1 and S_2 (i.e., the hollow ball $S^2 \times [1,2]$), and let b, m, w be the line segments in which H intersects the positive coordinate axes x, y, z, respectively. In a conversation Jože Vrabec asked John Milnor the following question:

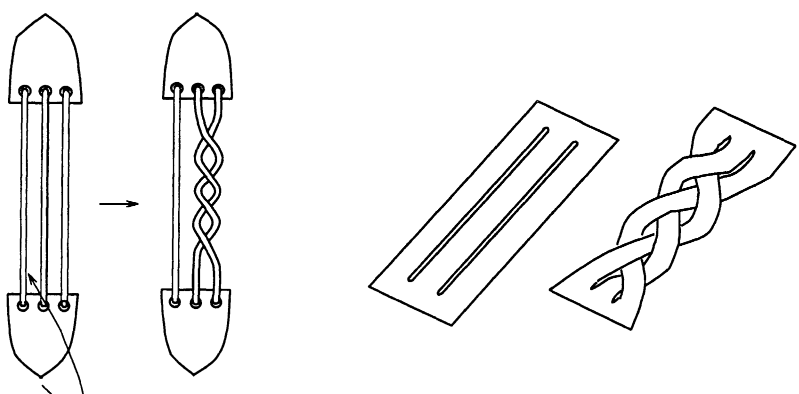
Question 1. *Is there an ambient isotopy of the hollow ball H fixing the spheres S_1 and S_2 that wraps m twice around w (see Figure 3(a))?*

We have already seen a graphical proof at the end of the previous section. In his Topology I (1985/86) lectures, Professor Vrabec demonstrated John Milnor's topological solution; he used the nontrivial element of the fundamental group of $SO(3)$ ($\cong RP^3$) to build the desired ambient isotopy, see Appendix or [**Br**, pp. 164–167]. Intriguingly, neither our graphical solution nor his solution can be easily used to answer the remaining part of our problem:

Question 2. *Is there an ambient isotopy of the hollow ball H fixing the spheres S_1 and S_2 that wraps m once around w?*

It turns out that both questions have a long history. E. D. Bolker's paper [**Bo**], which appeared over 20 years ago in the Monthly, gives another topological solution of Question 1. Physicists were, however, the first scientists concerned with Question 1. They used its solution in order to better understand the quantum nature of the electron. Swiss physicist W. Pauli used the calculus of spinors to model electrons. His model implies that the electron has to spin for 4π radians in

order to return to its starting position. In 1928, at the age of 26, P. A. M. Dirac described the movement of an electron with an equation that united *quantum mechanics* and the *special theory of relativity*. In order to convince his students that Pauli's model is not so surprising, he took a wrench (instead of our ball in the middle of the room) connected to a chair with three strings, rotated the wrench for 4π radians and then untangled the strings without moving the wrench. To learn more about the calculus of spinors and the theory of angular momentum, see [**BL**, Ch. 2]. In the early thirties this was such a hot topic at Niels Bohr's Institute for Theoretical Physics that some games were even invented. The Danish poet, writer, and mathematician Piet Hein called his game *Tangloids*, and it has been played for many years in Europe [**G**, p. 28]. Each of the two players holds one piece of wood with three holes that are connected by three shoe laces, see Figure 5(a). One player rotates her/his piece around *any* axis for 4π radians and the other one tries to untangle the strings while both pieces are allowed only to translate (rather than rotate). Then they reverse the roles and the one who untangles the shoe laces faster is the winner. (Why is this a good model for Question 1?) There is another related puzzle [**BCG**], in which you have to braid a paper (Figure 5(b)) without tearing it and using glue. The underlying principle is most probably quite old and known to craftsmen in leather (scouts should know about it as well). Hint: start braiding at one end and then untangle at the other end.



(a) (b)

Figure 5

By contrast, Question 2 concerns *nonexistence* and is therefore certainly in the domain of mathematics. It has been solved at least in two ways: Newman employed Artin's braid theory [**New**], cf. [**FB**], and Fadell used configuration spaces [**Fa**].

It seems that everything has already been done, except that there is no obvious connection between solutions of Question 1 and Question 2. One can still wonder why that graphical solution (Figure 4) works; checking it reminded us of a towing horse seeing only the path in front of itself. Could we understand it better or find a more natural solution, and use that to solve Question 2?* Let us use our

*At that time the author was not even aware that Question 2 had already been solved.

imagination (if not some childhood memories) and picture a girl/boy sitting at her/his desk, trying to satisfy the parents by studying some thick boring book. She/He soon gets tired and starts to play with a plastic model of an airplane. She/He wants to fly the plane forward on and on (planes do not fly backwards, remember?) while holding it firmly in one hand, however (s)he cannot move from the chair (the floor is squeaky and the parents might realize that (s)he is not studying any more), and (s)he wants to fly the plane without turning it upside down. She/He pretends to be a pilot of a commercial plane and passengers would get angry if (s)he did otherwise, at least those who do not use their seatbelts. Suddenly (s)he finds out that (s)he can move the plane along a path that has the shape of number 8. "Amazing", (s)he thinks, "if I moved the plane along a path in the shape of a circle my hand would tangle, and I would have to stop, but 8 works".

Could we use this ingenious solution in our problem? Well, let us twist and fold this figure of eight so that we see from the top only one "circle"; a mathematician would say that such an eight covers a circle (see Figure 6). The girl/boy has still no problem flying the plane in the path of this folded eight shape. Try it yourself; Figure 7 might be helpful (similar pictures have appeared in many journals and books, e.g., **[Ri]**, **[Str]**, **[Br**, p. 166]). But the girl/boy is now moving the plane in the path of a circle (for that reason we fold 8 completely), with the arm returning to the same position only after two full circles (during the first circle the elbow goes above the plane, while during the second circle it stays below the plane). There is no problem in making these circles so small that the plane is eventually just rotating. So the girl/boy can *rotate the plane in the same direction constantly while holding it firmly in the hand.* Actually, we can even say that (s)he can rotate the hand constantly in the same direction while it is turned up all the time. This phenomenon will be the leading idea of our proof of Question 2, and will contribute to finding a 'better' solution of Question 1. Are you puzzled? Think about it. If the plane were fixed, the hand would get twisted several times. So, in order to understand what is going on when we are moving elastic bands, we follow their twisting.

4. TWISTING NUMBER. Let us replace the line segments b, m, w with narrow bands. Such an approach is quite natural these days in knot theory and already has real-world applications. For example, when we study DNA molecules, which are long, thin, and usually double-stranded, we follow their twisting.

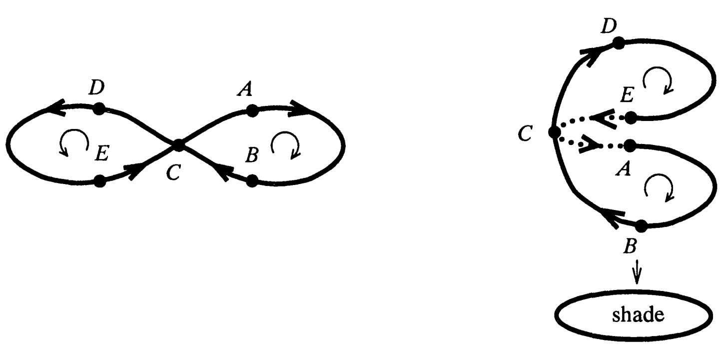


Figure 6

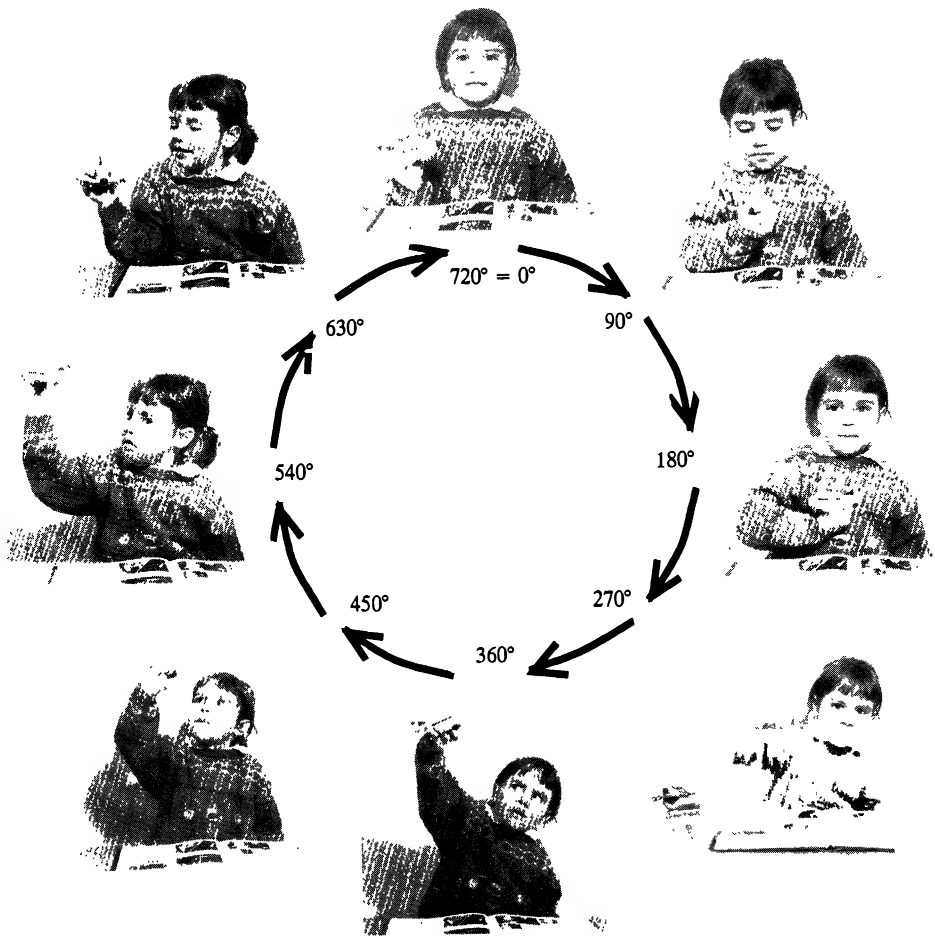


Figure 7. Eva Jurišić rotates plane while holding it firmly in her hand. For a similar demonstration by Uroš Jurišić, see http://crypto2.uwaterloo.ca:80/~ajurisic.

Before we attempt to solve Question 2 in this extended version (with bands) we study natural simplifications. The hand holding the plane could be considered as a single band. We try to find out what can happen to a single band in a three-dimensional ball (Lemma 1), and in a hollow ball H (Lemma 2). But first we introduce one of the simplest nontrivial invariants of ambient isotopy studied in knot theory. This is the linking number, defined for a pair of oriented curves in $\mathbb{R}^3$ (see Kauffman **[K1]** or **[K2]**). If a sign is associated to each crossing of the two-dimensional representation of the two curves as described in Figure 8, then the *linking number* is defined to be the sum of signs over all crossings of one closed curve with the other divided by two; signs of a curve crossing itself are not included. Since the

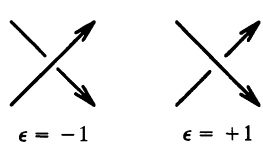


Figure 8. Right-hand-rule.

linking number is an ambient isotopy invariant, it does not depend on the two-dimensional representation we have chosen. In the nineteenth century, when knot theory was in its infancy, Gauss computed inductance (linking numbers) in a system of linked circular wires. The linking number plays an important role in the study of the effects of enzymes on a circular DNA, i.e., one that can be represented by a closed band in $\mathbb{R}^3$ whose boundary has two components [**Su1**].

Let X be a band of small width in a three dimensional subspace of $\mathbb{R}^3$ placed with its core (Figure 9(a)) along a line, with its ends in a plane, and both sides of the band oriented from the left end to the right end (Figure 9(b)). If a sign is associated again, as described in Figure 8, to each crossing of one side with the other in a two-dimensional representation in the plane that contains their ends, then the *twisting number*, $\mathrm{Tw}(X)$, is defined to be the sum of all signs divided by two. Intuitively, the twisting number is half the difference between the number of twists through the angle π and the number of twists through the angle $-\pi$.

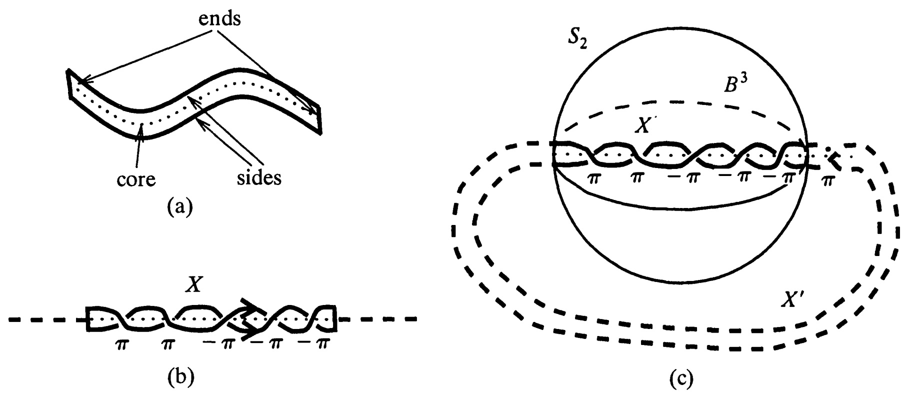
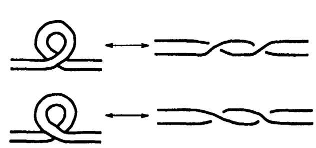


Figure 9

We could avoid relying on a two-dimensional representation, and give White's more general definition of this quantity by a line integral, which sums the amount of twisting along the core [**Su2**, p. 22], [CS]. However it would still not be an ambient isotopy invariant. Further, if our band were a closed band in three-dimensional space, then we could also define a writhing number, $\mathrm{Wr}(X)$, by orienting the core, assigning, as in the case of the linking number, a sign to each crossing, and then summing up all the signs. Then the sum of the writhing number $\mathrm{Wr}(X)$ and the twisting number $\mathrm{Tw}(X)$ equals the linking number $\mathrm{Lk}(X)$. This is J. H. White's Conservation Law [W]: $\mathrm{Lk}(X) = \mathrm{Wr}(X) + \mathrm{Tw}(X)$; see Figure 10.



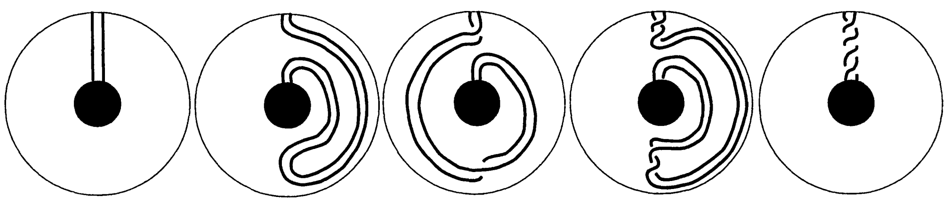
one writhe = one twist

Figure 10

Lemma 1. *Let B^3 be a ball in $\mathbb{R}^3$ that has S_2 for its boundary. If X is a band of small width, placed along x-axis, and with its ends on the equator, then the twisting number of X cannot be altered by an ambient isotopy of B^3 fixing S_2.*

Proof: We join the ends of the band X in the exterior of B^3 with a band X' so that the boundary of $X \cup X'$ has two components; see Figure 9(c). Any ambient isotopy of B^3 fixing S_2 can be extended with the identity on the complement of B^3 to an ambient isotopy of $\mathbb{R}^3$. If the ambient isotopy of B^3 changed the twisting number of X, then the ambient isotopy of $\mathbb{R}^3$ would change the linking number of the boundary of $X \cup X'$, which is impossible. □

Hence, when we examine the twisting number of a band in a three-dimensional ball it suffices to follow only the sides of the band. This is not the case with a hollow ball. Let A be a band of small width from S_1 to S_2 in H, placed along z-axis and with its ends in the xz plane. There exists an ambient isotopy of H that alters the twisting number of A by two. This should already be clear from the girl/boy's game with the airplane, but it is even more transparent from Figure 11.



1 2 3 4 5

Figure 11. We get from the second figure to the third one by flipping the part of the band closer to the sphere S_2 in the upper hemisphere. The next step is flipping the same part of the band in the lower hemisphere.

The last position can be drawn in another way, see Figure 12.

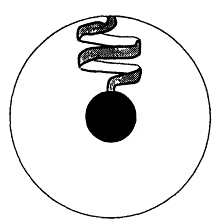


Figure 12

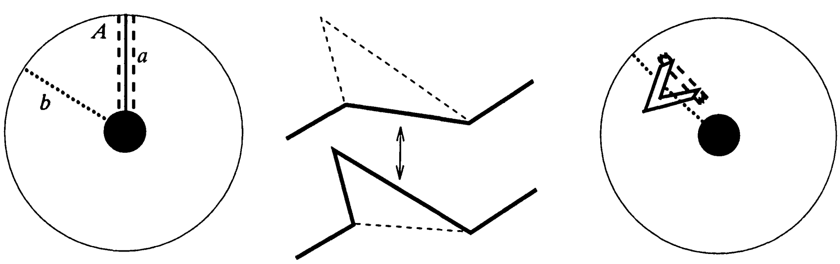
We now come to the heart of our problem. We are in a situation where we have to use a little bit of topology to prove the nonexistence part. However, due to the following two facts, we can stick with a combinatorial approach.

(1) Since our starting position can be drawn with polygonal lines, any position can be drawn with polygonal lines after a small deformation.
(2) If there is an ambient isotopy between two polygonal positions, then there exists a piecewise linear ambient isotopy between these positions [**BZ**, p. 4 and Corollary 3.16].

Lemma 2. *An ambient isotopy of H fixing S_1 and S_2 can alter the twisting number of the band A only by a multiple of two.*

There is evidently no ambient isotopy of the hollow ball H that alters the twisting number of the band A by one half, since the spheres S_1, S_2, and therefore also the ends of A, are fixed. The remaining part, namely that the twisting number cannot be altered by one, appears implicitly in [**K1**, VI.1, VI.18] without a proof; see the end of this section. We sketch our proof:

Proof: Let us denote by a and b the line segments from $(0,0,2)$ to $(0,0,1)$ and from $(2,0,0)$ to $(1,0,0)$ respectively (Figure 13(a)). So a is the core of the band A and we assume that it moves with an ambient isotopy of H. On the other hand, b will be considered as a ray, which does not move. As previously noted, we can assume [**BZ**, Prop. 1.10] that an ambient isotopy is piecewise linear, which implies that the induced movement of a is realized by a finite number of Δ-moves (Figure 13(b)) and that after every such move A and b are disjoint. It is important to note that as long as the band A does not go through the ray b, its twisting number cannot be changed, by Lemma 1. Let us now see what happens when during some Δ-move the band A goes through the ray b. We will show that we can move a to the same place, without going through the ray b, but at the cost of making two new twists of the band A; see Figure 14, which is obtained by slightly modifying Figure 11. Unfortunately we are not sure that we can perform the 'move' from Figure 14 in the hollow ball H, since some parts of the band A can be in the way. However, we can get rid of these difficulties, since we actually do not need the whole hollow ball H for that move. The part of the band on which we performed the Δ-move when taking the band A through b, lies at the outset in some hollow ball $H' \subset H$; see Figure 15(a). Assume that this hollow ball H' is also being deformed by our ambient isotopy; see Figure 15(b). But no matter how deformed, it will never



(a) (b) Delta move (c)

Figure 13

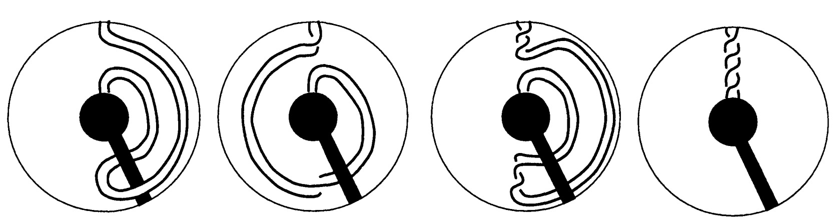
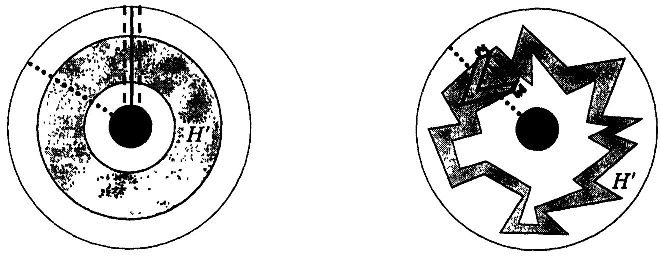


Figure 14

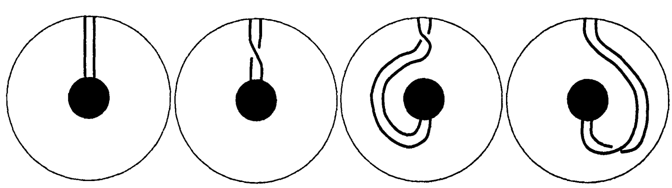


(a) (b)

Figure 15. Two plane sections.

contain any other part of the band A and it will always look like (i.e., be homeomorphic to) a hollow ball. So now we use this 'star-like' hollow ball H' to perform the move from Figure 14 backwards. Finally we can conclude that the twisting number of the band A can change only by a multiple of two. □

A reader who is anxious to see the solution of Question 2 can now proceed to the next section. But on the way let us mention an interesting connection with the quaternion group (see **[K1]** and **[K3]**). Let i, j, and k be the rotations of S_1 around the coordinate axes x, y, and z, respectively, through π radians. Let us define a group generated by i, j, k modulo the equivalence relation $\sim$ of an ambient isotopy in a hollow ball H with the band A keeping S_1 and S_2 fixed. Then Figure 11 shows that $i^4 = j^4 = k^4 = 1$. Similar pictures show that $i^2 = j^2 = k^2$, $ij = k$, $jk = i$, $ki = j$ and if we denote i^2 with minus one (-1) then also $ji = -k$, $kj = -i$, $ik = -j$. Figure 16 proves the last equality. Thus this group is exactly the quaternion group.



i ik $j^{-1} = j^3 = -j$

Figure 16

The nonexistence part of Lemma 2 is expressed in **[K1]** as $i^2 \neq 1$.

5. SOLUTION. We are now ready to apply Lemma 1 and Lemma 2 to Question 2. Let B, M, W be untwisted bands in H along line segments b, m, and w, respectively. Suppose there exists an ambient isotopy of H fixing S_1 and S_2 that wraps m k times around w, for some $k \in \mathbb{Z}$. If the width of the bands is small enough then this isotopy also wraps the band M k times around the band W and we have the position shown in Figure 17. If we orient the sides of the bands as in Figure 17, then, by Lemma 2, the twisting number of B is equal to $2n$, for some

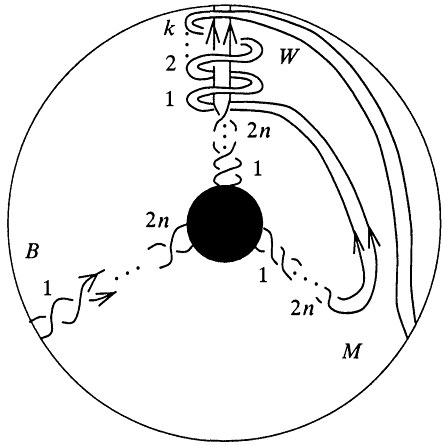


Figure 17

$n \in \mathbb{Z}$. Look at $B \cup W$. We can connect B and W by an untwisted band lying in the interior ball to form a single band with its ends on the boundary of the big ball. Then it follows from Lemma 1 that the twisting number of W is $-2n$. A similar argument can be applied to $B \cup M$: since we can pull M straight (moving it through W, of course) without introducing any new twists, the straight part of M has $-2n$ twists. On the other hand, when dealing with $M \cup W$ we are not allowed to move M through W (or itself or the interior ball). In order to pull M straight we first get rid of coils of M by using the move described in Figure 14 exactly k times (moving M through B, of course). Pulling M straight introduces $2k$ additional twists in M. Therefore the twisting number of M equals $-2n + 2k$. But on the other hand, by Lemma 1 for $M \cup W$, it equals $2n$. Hence $k = 2n$. Therefore there is no ambient isotopy of H fixing S_1 and S_2 that wraps M an odd number of times around W.

6. CONCLUSION. If we apply the result of Question 2 to the core and the sides of the band A in H, we get the nonexistence part of Lemma 2. Therefore we conclude that the result of Question 2 and the nonexistence part of Lemma 2 are equivalent.

Let us now give a more natural solution of Question 1. First we can move the elastics from the starting position in Figure 3(a) to the position in Figure 18. Second, we repeat the move from Figure 11 for the parallel parts of m and w in order to wrap m twice around w. Now it remains to untangle b with the "parallel" parts of m and w (remember that we can always untangle two elastic strings from S_1 to S_2).

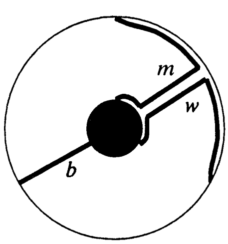


Figure 18

Finally we explain the phenomenon mentioned at the beginning. Suppose you coiled a long extension cord around your left shoulder. Imagine a three-dimensional ball in place of your shoulder, with the North Pole pointing in the direction of your left hand. So the cord gets coiled along the equator from the direction of the North Pole; see Figure 19(a). After the cord is stored for some time, we usually do not remember where the North pole is, and we unwind it either in the direction of the North or the South Pole. In the first case the cord will not be twisted, while in the second case (when we start to uncoil the cord straight from the wall where we stored it, see Figure 19(b)) Figure 11 guarantees that it will be twisted many times. Note that if you perform this experiment the cord either has to be long enough or you should fix one end and hold the other one with your hand. Lawn mowers, who are aware of the twisting problem, coil their extension cords in the shape of number 8, cf. Figure 6 and Figure 10. In industry, on the other hand, long extension cords are coiled on rolling cylinders.

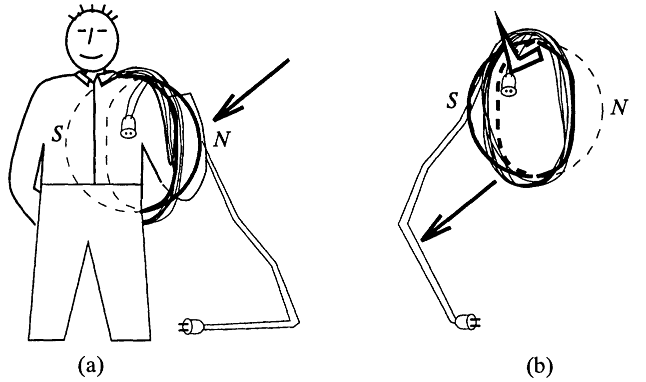


Figure 19

Let us conclude with applications from [**Sto**]. Figure 11 can be redrawn one more time (see Figure 20) in order to demonstrate that when the cord (in the shape of a question mark) goes around the small sphere once and the small sphere is rotated through 4π radians in the same direction, the cord resumes its original position. D. A. Adams used this fact to build a device that can supply electric power to a rotating platform through flexible wires (i.e., one that prevents the wires from twisting up and breaking). The reverse approach has actually been used widely in the electrical manufacturing industry, for example, to twist electrical wires uniformly. Many similar but simpler machines were discovered later by physicists [**Ri**].

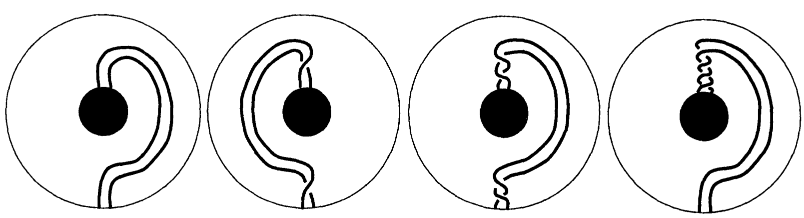


Figure 20

We can push the discrete approach further. Suppose we want to record a specimen on a rotating platform, but we do not want the viewer of our recording to notice the rotation. Can you find a solution where the camera does not move around? (Hint: Replace the band in Figure 20 with a few prisms.) A solution is shown in Figure 21. The same principle is used in a *periscope*, i.e., an instrument by which an observer obtains an otherwise obstructed view, to be able to look around while not moving her/his head; see [**BW**, pp. 243–244].

We hope that some teachers will find useful material for entertaining students while introducing the quaternions, for example. For further reading on related subjects see [**Re**], [**K1**], [**K2**], [**K3**], [**Br**], [**BZ**], [**Sti**] for knot theory, [**Su1**], [**Su2**] for applications in studies of DNA, and [**W**], [**CS**] for theoretical background on linking, twisting, writhing, and winding numbers.

APPENDIX. A topological solution of the first question is very well known to many topologists and can be found in many places, e.g., [**Br**]. Here we present the solution outlined by Vrabec.

This time we assume that the elastic bands b and w are along the x-axis. Let us define a loop u: $[0,1] \to SO(3)$ by

$$u(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 2\pi t & -\sin 2\pi t \\ 0 & \sin 2\pi t & \cos 2\pi t \end{bmatrix};$$

we can think of u as a 'continuous' rotation of $\mathbb{R}^3$ for 2π radians around x-axis. The concatenation $u * u$ is then a rotation through 4π radians. The action of the set $\{(u * u)(t)\colon 0 \le t \le 1\}$ on $P = (0,1,0)$ gives us the mapping (loop) $[0,1] \to S^2$, $t \mapsto [(u * u)(t)](P)$. The graph of this mapping is a path in $[0,1] \times S^2$. If we identify this product with the hollow ball in a natural way, the graph of the mapping mentioned above corresponds to a path that is wrapped twice around the x-axis. It is well known that $SO(3) \cong RP^3$ and so $\pi_1(SO(3)) = \pi_1(RP^3) = \mathbb{Z}_2$. Therefore there exists a homotopy between $u * u$ and $[0,1] \to \{\mathrm{id}\} \subset SO(3)$; the equivalence class of u is actually the only nontrivial element of the fundamental group of $SO(3)$. So let

$$\{F_s\colon [0,1] \to SO(3) : 0 \le s \le 1\}$$

be the corresponding homotopy, $F_0(t) = \mathrm{id}$, and $F_1(t) = (u * u)(t)$ for each t. From this we get the ambient isotopy (family of homeomorphisms)

$$\{G_s\colon [0,1] \times S^2 \to [0,1] \times S^2 : 0 \le s \le 1\}$$

of the hollow ball $[0,1] \times S^2$ in the following way: $G_s(t, Y) := (t, [F_s(t)](Y))$ for $t \in [0,1]$, $Y \in S^2$. Evidently we have $G_0 = \mathrm{id}$ and G_1 is the mapping that maps $[0,1] \times \{P\}$ to the path mentioned above that is wrapped twice around the x-axis. The ambient isotopy G pointwise fixes the boundary spheres, since $G_s(i, Y) = (i, Y)$ for $i \in \{0,1\}$ and $Y \in S^2$, i.e., $F_s(i) = I_3$ (the identity matrix 3×3), which follows from the fact that F_s is the homotopy of the loops with starting and ending points at $\mathrm{id} \in SO(3)$. Finally $G_1(t, (x,0,0)) = (t, (x,0,0))$, since $F_1(t) = (u * u)(t)$ is identity at $(x,0,0)$ (i.e., rotations around the x-axis fix points on the x-axis), so after homotopy is applied b and w are at their starting places. □

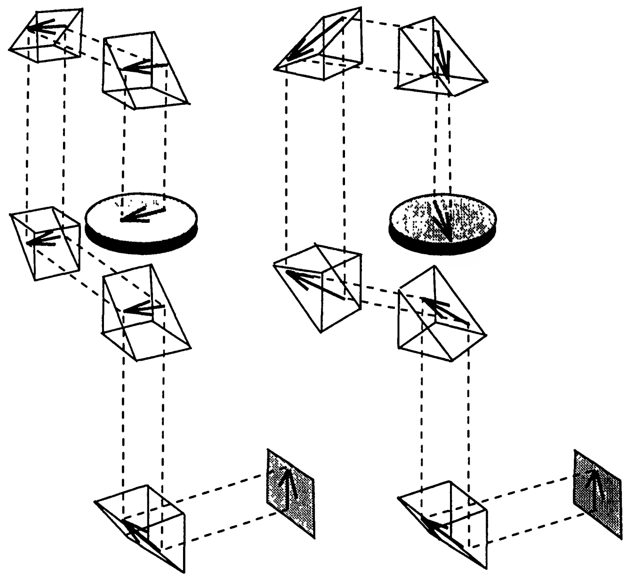


Figure 21. Train of prisms 'untwists' an optical beam.

ACKNOWLEDGMENTS. I explored some of the ideas in this paper in conversation or correspondence with W. J. Gilbert, F. Jaeger, U. Milutinović, R. C. Read, and J. Vrabec. I also thank several colleagues who read earlier versions of my manuscript.

REFERENCES

[BCG] E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning Ways for Your Mathematical Plays.* Academic Press, London-Toronto, 1982.

[BL] L. C. Biedenharn and J. D. Louck, Angular Momentum in Quantum Physics, Theory and Application, in *Encyclopedia of Mathematics and its Applications*, Addison-Wesley, 1981.

[Bo] E. D. Bolker, The spinor spanner, *Amer. Math. Monthly* **80** (1973), 997.

[BW] M. Born and E. Wolf, *Principles of Optics*, Pergamon, Oxford, 1970.

[Br] G. E. Bredon, *Topology and Geometry.* Springer-Verlag, New York-Berlin-Heidelberg, 1993.

[BZ] G. Burde and H. Zieschang, *Knots.* De Gruyter Studies in Mathematics **5**, Berlin-New York, 1985.

[CS] W. G. Chinn and N. Steenrod, *First concepts of topology.* Mathematical Association of America, Washington, DC, 1966.

[Fa] E. Fadell, Homotopy groups of configuration spaces and the string problem of Dirac, *Duke Math. J.* **29** (1962), 231–242.

[FB] E. Fadell and J. van Buskirk, The braid groups of E^2 and S^2, *Duke Math. J.* **29** (1962), 243–257.

[G] M. Gardner, *New Mathematical Diversions from Scientific American.* Simon and Schuster, New York, 1966.

[K1] L. H. Kauffman, On Knots. *Annals of Mathematics Studies*, Princeton University Press, Princeton, 1987.

[K2] L. H. Kauffman, New Invariants in the Theory of Knots, *Amer. Math. Monthly* **95** (1988), 195–242.

[K3] L. H. Kauffman, *Knots and Physics.* World Scientific Pub., Singapore-New Jersey-London-Hong Kong, 1991.

[New] M. H. A. Newman, On a string problem of Dirac, *J. London Math. Soc.* **17** (1942), 173–177.

[Neu] L. Neuwirth, The theory of knots, *Scientific American*, June, 1979, pp. 84–96.

[Re] K. Reidemeister, *Knotentheorie.* Chelsea Publ. Co., New York, 1948; copyright 1932, Julius Springer, Berlin.

[Ri] E. Rieflin, Some mechanisms related to Dirac's strings, *Amer. J. Physics* **47** (1979), 379–381.

[Sti] J. Stillwell, *Classical Topology and Combinatorial Group Theory*. Springer-Verlag, New York-Berlin-Heidelberg, 1993.

[Sto] C. L. Stong, *The Amateur Scientist* (Subtitle: Diverse topics, starting with how to supply electric power to something that is turning), *Scientific American*, December, 1975, pp. 120–125.

[Str] J. Strnad, Še o sukanju, Presek **19** (1991/92), 320–321 (in Slovene).

[Su1] D. W. Sumners, Untangling DNA, *Math. Intelligencer* **12** (1990), 71–80.

[Su2] D. L. Sumners, editor, *New Scientific Applications of Geometry and Topology*, Proceedings of Symposia in Applied Mathematics, vol. 45, American Mathematical Society, Providence, 1992.

[W] J. White, Self-linking and Gauss integral in higher dimensions, *Amer. J. Math.* **91** (1969), 693–728.

Department of Combinatorics and Optimization
University of Waterloo
Waterloo, Canada, N2L 3G1
ajurisic@crypto2.uwaterloo.ca

It's De-Lovely

(With apologies to Cole Porter)

You take two sets, you form their meet,
Complementation is quite a treat,
It's delightful, It's DeMorgan, it's d'algebra

You grab a 'zee' in the complex plane,
Taking its power is not a pain,
It's delicious, it's DeMoivre, it's d'analysis.

We can tell at a glance
Why the math folks sing "Vive la France ,
We can swear their view of logic's
Gotta be true, "true as l'bleu".

But don't take sides, just to have some fun,
All mathematics is one-to-one,
It' delightful, it's delicious,
It's de-stributive, it's dee-'wye'-dee-'ex',
It's d'limit, it's de-luxe, it's de-lovely.

Contributed by Ronald E. Praher, Trinity University

Fabian Stedman: The First Group Theorist?

Arthur T. White

1. INTRODUCTION. Fabian Stedman, a son of the Reverend Francis Stedman, vicar of Yarkhill, Herefordshire, England, was baptized there on December 7, 1640. At age fifteen he was apprenticed to the Master Printer Daniel Pakeman in London. In London he joined the Scholars of Cheapside, a bell-ringing society, serving as its Treasurer in 1662. The following year Stedman became a Freeman of the Stationers Company. In 1664 he joined the Society of Colledg Youths, which had been founded in 1637; renamed the Ancient Society of College Youths in the nineteenth century, this bell-ringing society is still active today. There is some evidence that Stedman moved from London to Cambridge in 1664 (see [**3**] and [**5**], which are the sources for much of this background information), and he might have been working as a printer there and also serving as parish clerk of St. Bene't's. The early 11th-century Saxon tower of St. Benedict's Church is the oldest surviving building in Cambridgeshire.

In 1677 Stedman became Steward of College Youths; five years later he was Master of the Society. Returning to (or staying in) London, he changed profession, becoming a clerk in the office of Audit of Excise. He died in 1713, and was buried at St. Andrew Undershaft on November 16.

Fabian Stedman's claim to fame as at least one of the "fathers of bell ringing" seems beyond doubt; his contributions to the first two books on change ringing, *Tintinnalogia* (1668) [**2**] and *Campanalogia* (1677) [**7**], will be summarized shortly. What is less well known, and what has occasioned this article, is the group theory latent in his writings and in his compositions—a full century before Lagrange wrote "Reflexions" (1770).

2. CHANGE RINGING. In England church bells are rung not in melody, but in permutations (changes). To a limited extent, this practice has spread to Australia, to Canada, and to the United States. The increase in control facilitated by the mounting of each bell on a circular wheel allowed the inception of change ringing in about 1610. Early forms of change ringing involved one row (one ordering of the bells, denoted by $1, 2, \ldots, n$; here $n = 6$), such as *rounds* (123456), *queens* (135246), or *tittums* (142536) to be rung repeatedly until the conductor (one of the ringers) called for a change; these are known as *call changes*. Due to mechanical considerations arising from the manner of mounting the bells, if the rows are changed constantly then no bell can readily change its order of striking by more than one position. Thus each *change* (a transition from one row to the next) involves one or more disjoint pairs of adjacent bells swapping over. At first *plain changes*, involving one pair only at each step, were in vogue. The four rows in Figure 1a illustrate three successive plain changes on six bells, commencing with rounds. Soon, *cross changes*, allowing more than one swapping pair, replaced plain changes, continuing to the present day. As we will see, Stedman was instrumental in effecting this

(a)	(b)
123456	
213456	123456
214356	214365
214365	

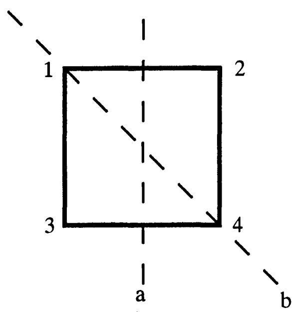
Figure 1. Three Plain Changes and One Cross Change.

transition. In Figure 1b we see how to get from rounds to the last row in Figure 1a by using one cross change, instead of three plain changes.

In about 1621, cross and plain changes were alternated to produce the plain lead on four bells, as shown in Figure 2a. From a modern viewpoint, if we let $a = (12)(34)$ denote the cross change that swaps both the first two and the last two bells and $b = (23)$ the plain change that swaps the middle pair, and if we note that reflections a and b generate the dihedral group D_4 (as the group of symmetries of a square labelled as in Figure 2b), then we see that the eight rows of Figure 2a coincide with the elements of D_4. This lead (it could also be called the *hunting group* [**12**], as the *treble*—bell 1—is *plain hunting* in this group of rows) is described by the identity word $(ab)^4 = e$ in the symmetric group S_4. In change ringing, every *touch* (on n bells, say) begins and ends with rounds, and thus is described, in modern terms, by an identity word in S_n, where each letter of the word is an involution in S_n consisting of disjoint pairs of adjacent interchanges.

1234
2143
2413
4231
4321
3412
3142
1324
1234

(a)



(b)

Figure 2. The Plain Lead on Four Bells.

3. THE TWO MAIN SOURCES. *Tintinnalogia* [2] was published in 1668, written "By a Lover of that ART", and "printed by W. G. for Fabian Stedman, at his shop in St. Dunstans Churchyard in Fleetstreet". It is thought [3] that "W. G." stands for the publisher W. Godbid, that Stedman helped to arrange the printing and perhaps helped to supply material for the book, but that the actual author was Richard Duckworth. Duckworth was a fellow of Brasenose College, Oxford and later rector of St. Martin's, Carfax, Oxford. Fabian Stedman's father also was a member of Brasenose College and Fabian's older brother Francis was contemporaneous with Duckworth at Oxford. This might explain the connection between Fabian Stedman of London/Cambridge and Richard Duckworth of Oxford.

At any rate, in *Tintinnalogia* its author describes first plain changes and then, tentatively, cross changes, extending Figure 2a to its full extent, 24 rows now

known as *Plain Bob Minimus* (discussed below), and goes on to discuss peals on five and six bells. It is notable that leads of various peals are written out in full, with no omitted rows. In 1671 a Second Edition of *Tintinnalogia* was printed ("for F. S."); only one copy of the Second Edition survives, at the Bodleian Library, Oxford. This was not an updating of the first edition, but a reprinting.

The updating occurred in 1677, with the publication of *Campanalogia* [7], printed "by W. Godbid, for F. S.", and it was substantial. The Epistle Dedicatory for this work, to the "Society of COLLEDG YOUTHS", is signed "A constant Well-wisher to the Prosperity (though an unworthy member) of your Society, F. S." As the College Youths' name book for this period shows only one name with initials F. S.—Fabian Stedman—(see [4]), it seems safe to attribute authorship of *Campanalogia* to him. At the outset (page 2) Stedman says

> Although the practick part of *Ringing* is chiefly the subject of this Discourse, yet first I will speak something of the Art of *Changes*, its Invention being Mathematical, and produceth incredible effects, as hereafter will appear.

In the Epistle Dedicatory, Stedman had referred to the plain lead on five bells (ten rows, generated from rounds by $(ab)^5 = e$, where $a = (12)(34)$ and $b = (23)(45)$, analogous to Figure 2(a)) as follows:

> ... it was thought impossible that double changes on five bells could be made to extend further than *ten*...

The blockage was evidently caused by the awareness that the two changes generated a closed system, what we now call the dihedral group D_5, and was relieved by the discovery that one closed system can be enlarged to another by the addition of new elements (in this case adding first $c = (34)$ to produce the *plain course* $[(ab)^4ac]^4 = e$ and then the bob $d = (45)$ to ring all of Plain Bob Doubles: $[((ab)^4ac)^3(ab)^4ad]^3 = e$, as on page 104 of *Campanalogia*—except that Stedman calls the bob an *extream*, and refers to the composition as *Old Doubles*).

After giving lengthy instruction on factorials and discussing the "Practice of Ringing" and plain changes, Stedman describes a number of cross peals coinciding with those in *Tintinnalogia*, with the innovation that only the first two leads are written out in full (a *lead* is a block of rows, such as the first eight in Figure 2a, from one *treble lead*—bell 1 in the first position—to the next); subsequent leads were represented by only their first and last row (both treble leads), as the rows between can be reconstructed from the pattern of the first leads given. The crucial point here is that a subset of rows is being represented by two of its elements. As the last row of a lead can be reconstructed from the first row, the last row is also superfluous to list, but Stedman continued to do so in order to make more readily apparent the change used to get to the first row of the next lead.

The second half of *Campanalogia* contains a large number of new methods, on five, six, seven, and eight bells, including fifty-three of Stedman's own compositions under the heading "London Peals". Other venues represented are Nottingham, Oxford, and Cambridge. Included among the fifty-three London peals is "Stedman's Principle", now known as "Stedman Doubles". Extendable to any odd number of bells (and even numbers as well, by having the tenor (bell n) ring last (*in cover*) in every row), this composition is one of the most popular to this day.

4. IMPLICIT ELEMENTS OF GROUP THEORY. To make the point that group theory is latent in change ringing to a substantial degree, we next analyze two

1234	1342	1423
2143	3124	4132
2413	3214	4312
4231	2341	3421
4321	2431	3241
3412	4213	2314
3142	4123	2134
1324	1432	1243
		1234

Figure 3. Plain Bob Minimus.

compositions—Plain Bob Minimus and Stedman Doubles—in some detail. We then discuss the extent to which Fabian Stedman was aware of these connections.

In Figure 3 we list all twenty four rows of Plain Bob Minimus in three columns; each column gives one lead of the full extent. At the end of the third column, we show the required return to rounds. As for all extents (which ring the full $n!$, on n bells), it is crucial that no other row is repeated.

As before, let $a = (12)(34)$ and $b = (23)$ describe possible changes from one row to the next. Add $c = (34)$ and let e denote the identity element of S_4. As before, a and b generate the dihedral subgroup D_4 of S_4, and the rows of the first lead correspond to $D_4 = \{e, a, ab, aba, (ab)^2, (ab)^2a, (ab)^3, (ab)^3a\}$. (From a modern point of view—see [**8–11**], for example—$ab = (12)(34)(23) = (1243)$, composing right to left, which we interpret as ringing in position 1 bell 2, in position 2 bell 4, in position 4 bell 3, and in position 3 bell 1; that is, the row 2 4 1 3. This extends in a natural manner to a full correspondence between rows and permutations.) If we follow row $(ab)^3a$ by change b we would regain rounds prematurely, since $(ab)^4 = e$. Thus we employ change $c = (34)$ for the first time, obtaining the second column $\{w, wa, wab, waba, w(ab)^2, w(ab)^2a, w(ab)^3, w(ab)^3a\}$, where $w = (ab)^3ac$. But this is just the left coset wD_4! Using c a second time, we get the third column as the final left coset w^2D_4, and a third and final use of c returns us to rounds.

A composition such as Plain Bob Minimus is required to satisfy six conditions, which we now list, together with an algebraic verification for each one.

(i) The extent must begin and end with rounds. (This follows from $[(ab)^3ac]^3 = e$.)
(ii) No other row is repeated. (The coset decomposition guarantees this.)
(iii) From one row to the next, no bell moves more than one position. (This is forced by our choice of $a = (12)(34)$, $b = (23)$, and $c = (34)$.)
(iv) No bell rests in the same place for more than two successive rows. (The alternation of $a = (12)(34)$ moves every bell appropriately.)
(v) The working bells (here, all but the treble) should all do the same work. (This is guaranteed by $w = (234)$, so that what bell 2 does in the first lead, bell 3 does in the second and bell 4 does in the third, etc.)
(vi) Each lead should be palindromic in its changes. (Examine $(ab)^3a$.)

These *axioms* for a *method*, as it is called, are not formally combined by Stedman in *Campanalogia*. His "Obser. 4" (pp. 38–39) corresponds to (iii); his other "observations" apply to performance, rather than composition. Axioms (ii)

and (v) appear in *Campanalogia* on pages 3 and 84 respectively. Axiom (i) is implicit throughout; (iv) and (iv) are more commonly employed in modern times. Modern ringers have more formally set forth these and other requirements (see, for example, [**1**, p. 8] and [**16**]), in what we now recognize as an axiomatic approach.

In summary, the decomposition of S_4 into left cosets of D_4 shown by Figure 3 precisely describes Plain Bob Minimus. (In *Campanalogia* (p. 96) Stedman lists the 24 rows in one column, but he uses letters to show where each block of eight rows (i.e., each coset) changes into the next.) But there are at least four other coset decompositions of interest here. To describe these, it is helpful to think of Figure 3 as an 8×3 matrix, whose entries are the rows of Plain Bob Minimus. In what comes immediately below, the term *row** will refer to a row of this matrix, which consists of three rows of the composition.

(1) The rows* of the matrix are the right cosets of the subgroup $\{e, w, w^2\}$ of S_4.

(2) Rows* 1 and 8 give the subgroup $(S_4)_1 \cong S_3$ of S_4; the set of all treble leads is just the stabilizer in S_4 of (bell) 1. The other right cosets consist of rows* 2 and 7, rows* 3 and 6, and rows* 4 and 5. Note that the row* numbers of each coset are symmetrical about the half lead, and that each half lead constitutes a right transversal of $(S_4)_1$ in S_4 (each of the right cosets is represented exactly once, in each half lead). Note also that each element of the ith coset fixes bell 1 in the ith position, $i = 1, 2, 3, 4$ (in accordance with the plain hunt). These follow from the fact that $(ab)^3a$ is a palindrome (condition (vi)) and are useful in "proving" extents, as we do for Plain Bob Minimus below.

(3) Rows* 1, 2, 5, and 6 give the subgroup A_4 of S_4, consisting of all the even (*in- course*) rows. The other four rows*, which form the other (right or left) coset of A_4 in S_4, consist of all the odd (*out-of-course*) rows.

I believe that Stedman made use of all these decompositions, although of course he lacked the modern terminology for them. Here is one decomposition that was probably not used by Stedman.

(4) In Figure 2 of [**8**], a Cayley color graph $C_\Delta(S_4)$, with $\Delta = \{a, b, c\}$, is shown imbedded in the projective plane with 4-fold symmetry, as generated by $ab = (1243)$. The six right cosets of the corresponding subgroup allow an even simpler depiction of Plain Bob Minimus, as a Schreier coset graph. This idea has been exploited to great advantage in [**9**], [**10**], and [**11**].

Now we turn our attention to Stedman Doubles. The plain course (consisting of 60 rows) is given in Figure 4. For convenience the presentation differs slightly from that given by Stedman in *Campanalogia* (pp. 129–132); both differ from that used by ringers today. But the connection with group theory is unaffected. Letting $a = (12)(45)$, $b = (23)(45)$, and $c = (12)(34)$, we can describe the plain course by $w^5 = e$, where $w = (ab)^2ac(ba)^2bc = (13452)$. The sequence (word) ababa of changes gives a *slow six*; the sequence babab, which is used in alternation, gives a *quick six*. Each yields all the permutations on the front three bells, and thus a subgroup isomorphic to S_3, if we start with rounds. We introduce $c = (12)(34)$, called by Stedman a *parting change*, to link successive sixes—by exchanging one of the back two bells with one of the front three. Stedman notes: "Bt this method the peal will go sixty changes, and to carry it farther *extremes* must be made." We check that, with all changes (a, b, and c) even, the largest subgroup of S_5 we can generate is A_5. Figure 4 displays A_5 decomposed into ten left cosets of the subgroup isomorphic to S_3 given by the first six rows. Or, if we focus on the rows* of the 12×5 matrix, we find twelve right cosets of the subgroup generated by

12345	31452	43521	54213	25134
21354	13425	34512	45231	52143
23145	14352	35421	42513	51234
32154	41325	53412	24531	15243
31245	43152	54321	25413	12534
13254	34125	45312	52431	21543
31524	43215	54132	25341	12453
35142	42351	51423	23514	14235
53124	24315	15432	32541	41253
51342	23451	14523	35214	42135
15324	32415	41532	53241	24153
13542	34251	45123	52314	21435
				12345

Figure 4. Plain Course of Stedman Doubles.

$w = (13452)$, represented by the five rows of Stedman Doubles in the first row* of the matrix. All 60 rows are in-course. (We note in passing that Stedman Doubles, as a *principle*, has no hunt bell (bell 1 was plain hunting in Plain Bob Minimus, where bells 2, 3, and 4 were working alike); now all five bells are working alike, as forced by w being a 5-cycle. With this understanding, all "axioms" (i)–(iv) hold for this principle, just as they did for the method Plain Bob Minimus. However, the subgroup $(S_5)_1$ of S_5 plays no role here, and we have no coset decomposition to match (2) for Plain Bob Minimus.) To get the remaining 60 (out-of-course, i.e., odd) rows of Stedman Doubles, we replace the tenth use of the parting change c, which brought us back to rounds after 60 rows, by an appropriate change, say $d = (34)$—called by Stedman an *extream*, known now as a *single*. (The single (12) would also work here.) This throws us into the other coset of A_5 in S_5, and we get all of Stedman Doubles as $\{[(ab)^2ac(ba)^2bc]^4(ab)^2ac(ba)^2bd\}^2 = e$. Stedman's arrangement in *Campanalogia* (p. 131) clearly reflects this division of S_5 into cosets of A_5. However, he does not emphasize the division of A_5 (and the other coset) into cosets of S_3, as later change ringers have done.

5. STEDMAN A GROUP THEORIST? Certainly a knowledge of group theory helps us analyze (and compose!) pieces of change ringing music such as Plain Bob Minimus and Stedman Doubles for their structure and properties. Group theory, a mathematical discipline developed in the late eighteenth and nineteenth centuries, was of course not available to Fabian Stedman in 1677 and before, when he composed the music he recorded in *Campanalogia*. But was he in reality functioning as a very early group theorist in composing and verifying his compositions?

Of all the six requirements for change ringing given above, (ii) is by far the most difficult to verify: no row is repeated; each appears exactly once (except for rounds, which appears first and last, but nowhere else). The verification of (ii) is called *proof* by ringers, and if two rows that should differ in fact agree, then *falsity* has been established. Not all change ringing compositions correspond to left coset decompositions by a subgroup consisting of the rows of the first lead. But it is interesting to note that the two most popular methods (Plain Bob, on any even number of bells; Grandsire, on any odd number) and the most popular principle (Stedman, on any odd number) all do (except that the bob leads for Grandsire are not quite cosets). As mathematicians, we know that two left cosets are either disjoint or identical, and that no one coset has any internal falsity. Thus if we just

note that the three even (in-course) treble leads for Plain Bob Minimus (rows 1234, 1342, and 1423) are distinct, we have proved the composition.

But how might a ringer without explicit knowledge of group theory prove a composition like Plain Bob Minimus? Suppose, for example, the row 2341 appears twice. Since each lead is true, this must be in different leads. Since following 2341, which is out-of-course, by $a = (12)(34)$, $b = (23)$, and then a again gets us to an in-course treble lead (since row 2341 must be in either row* 4 or row* 5, and the palindromic condition (vi) guarantees that moving up by aba (row* 4) or down by aba (row* 5) will reach a treble lead head or a treble lead end respectively), the two leads containing 2341 must be headed by the same in-course treble lead. But a quick inspection shows that 1234, 1342, and 1423 are distinct. Thus 2341 *cannot* appear twice. A similar analysis applies to any other row. In summary, we need compare only one representative from each lead, even if we don't know that that lead is going to be called a coset more than a century later.

Did Stedman reason in this way? Here is what he said, on pages 94 and 95 of *Campanalogia*; for *whole hunt* read "treble;" for *course* read "lead;" for *peal* read "extent" (the full $n!$), for *pricking* read "writing."

> ... every note in a *cross-peal* must of necessity lie as many times in one place, as the rest of the notes are capable of making changes;

(In the 20th century, we would write $|(S_n)_i| = |S_{n-1}|$.)

> and also that two or more of the notes must jointly lie in the same places as many times, as the remaining number are also capable of making changes:

($|(S_n)_{i,j}| = |S_{n-2}|$, etc.)

> this being a certain touchstone to prove all *cross-peals* after they are prickt, and must be held as a principle upon which to ground such methods of pricking, that the course of all the notes may demonstrably tend to produce those effects. And from hence it is, that the whole *hunt* immediately derives the manner of its uniform motion through the courses of each peal. And the changes in every course are as so many guides to conduct the rest of the notes in such sort, that they may be prepared to lie at the last change of the course in apt places for each succeeding course to receive them, and to perform the like. Now as the changes in all the courses of a peal are made alike,... so in the composing of *cross-peals*, by pricking of one course may soon be discovered, whether a compleat peal will from these arise.

In connection with Stedman Doubles, the composer clearly seemed to know that following 60 true in-course rows (the first 59 "changes are all double", as he said on page 129 of [7]) by an appropriate "extream" would produce 60 true out-of-course rows, that $wx = wy$ means that also $x = y$. And, he seemed aware that the parity (in or out of course) of a row is dependent only on the row itself, not on its position in the composition. The modern theorem is that the parity (even or odd) of the number of transpositions into which a permutation can be decomposed is constant.

If we extend Stedman's principle on five bells, Stedman Doubles, to seven bells, we get Stedman Triples. Letting $a = (12)(45)(67)$, $b = (23)(45)(67)$, and $c = (12)(34)(56)$, we obtain the plain course $w^7 = e$, where $w = (ab)^2ac(ba)^2bc = (1374562)$; this plain course consists of 84 of the 5040 rows. To expand this touch

to a full extent, bob $d = (12)(34)(67)$ and single $f = (12)(34)$ have been used effectively, replacing c in either or both of its occurrences in certain subwords w, in order to get beyond the plain course, even as far as the full extent. Until recently, the most famous unsolved problem in bell ringing was the following: Is it possible to ring the full extent of Stedman Triples using only a, b, c, and d? In late 1994, Colin Wyld achieved such a composition [14], using, out of 840 positions where a bob might be called, 705 bobs [13]. Then, in early 1995, Andrew Johnson and Philip Saddleton also composed an extent of Stedman Triples using common bobs only (no singles), and one week later their composition was successfully rung by a Cambridge University Guild band, being called (579 bobs) at the first attempt by Philip Agg, at St. John's Waterloo Road [15]. See also Saddleton [6]. Thus a centuries-old (mathematical!) problem derived from the work of Fabian Stedman has finally been settled. The solution corresponds to a hamiltonian circuit in the Cayley graph for the symmetric group S_7, as generated by the involutions a, b, c, and d above, incorporating slow and quick sixes in alternation, linked by generators c and d.

I have not tried to make the case that Fabian Stedman was using group theory explicitly, but rather that group-theoretical ideas were implicit in his writings and compositions. These ideas, as we have seen, include closed systems, axiomatic systems, coset decomposition (including the ideas of coset representative and disjointness), even and odd permutations, factorials, and stabilizers in permutation groups. We should remember that those usually thought of as the first group theorists (Lagrange, Ruffini, Cauchy, Abel, and Galois) also were operating implicitly in the context of permutation groups, many decades before the definition of an abstract group as a set with a binary operation satisfying certain axioms.

REFERENCES

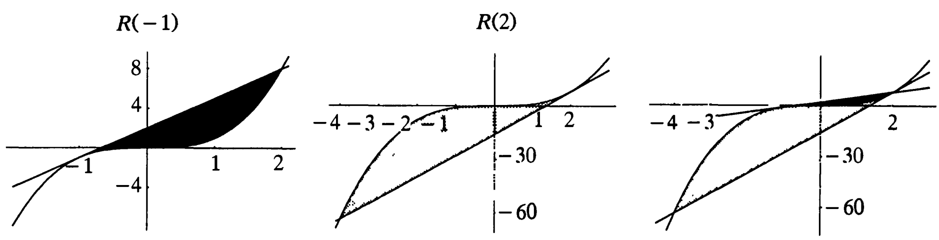
1. William T. Cook, Introduction to the art of change ringing, *Change Ringing: The History of an English Art*, J. Sanderson, ed., The Central Council of Church Bell Ringers, Cambridge, 1987, 1–17.
2. Richard Duckworth, *Tintinnalogia*, W. G., London, 1668.
3. John C. Eisel, Tintinnalogia, *Change Ringing: The History of an English Art*, J. Sanderson, ed., The Central Council of Church Bell Ringers, Cambridge, 1987, 50–58.
4. J. C. Eisel, A question of authorship, *The Ringing World*, Guildford, January 29, 1993, 93.
5. Ron Johnston, *Bell-ringing*, Viking, Harmondsworth, 1986.
6. Philip Saddleton, Key unexpectedly simple, *The Ringing World*, Guildford, March 3, 1995, 216.
7. Fabian Stedman, *Campanalogia*, W. Godbid, London, 1677.
8. A. T. White, Ringing the changes, *Math. Proc. Camb. Philos. Soc.* **94** (1983), 203–215.
9. A. T. White, Ringing the cosets, *Amer. Math. Monthly* **94** (1987), 721–746.
10. A. T. White, Ringing the cosets II, *Math. Proc. Camb. Philos. Soc.* **105** (1989), 53–65.
11. A. T. White, Treble dodging minor methods: ringing the cosets, on six bells, *Discrete Math.* **122** (1993), 307–323.
12. Arthur White and Robin Wilson, The hunting group, *Mathematical Gazette* **79** (1995), 5–16.
13. Colin Wyld, A 250-year old problem solved, *The Ringing World*, Guildford, February 24, 1995, 197–198.
14. "Around and About", *The Ringing World*, Guildford, January 27, 1995, 104.
15. "Editorial", *The Ringing World*, Guildford, February 3, 1995, 107.
16. *The Council's Decisions*, The Central Council of Church Bell Ringers, Cambridge, 1989.

Department of Mathematics and Statistics
Western Michigan University
Kalamazoo, MI 49008
white@wmich.edu

Functions Whose Successive Tangent Lines Enclose Proportional Areas

Herbert I. Brown, James W. Burgmeier, and David S. Dummit

1. INTRODUCTION. The tangent line to the curve $y = x^3$ at the point (a, a^3) with $a \neq 0$ intersects the curve in precisely one more point (with coordinates $(-2a, -8a^3)$), defining a region $R(a)$ between the curve and the portion of the tangent line connecting these two points. We can iterate this construction to define the region $R(-2a)$ and it is a perhaps surprising result that the area of this successive region is always precisely 16 times the area of $R(a)$, independent of the initial value of a. The figure below illustrates the successive regions $R(-1)$ and $R(2)$ for $a = -1$.



The fact that the ratio of these areas for $y = x^3$ is constant was determined experimentally by a student of the first author in the context of a classroom exercise (using MAPLE) to determine in terms of a the coordinates of the remaining point of intersection of the second tangent line in this construction.

The purpose of this paper is to consider several questions that arise from trying to generalize this example. We first make some general observations and prove that successive tangent lines enclose proportional areas for arbitrary cubic polynomials. The first generalization proves that a similar phenomenon occurs for arbitrary polynomials $f(x)$ of degree n when the tangent line is replaced by the Taylor polynomial approximation of degree $n - 2$. We then consider replacing x^3 by the extension of x^α to an odd function for positive $\alpha \neq 1$ and prove that the successive tangent lines again enclose proportional areas (these two possible generalizations coincide for cubics). The final two sections investigate the dependence of the proportionality constant on α and consider the converse question of determining all functions with the successive tangent property.

2. CUBIC POLYNOMIALS. Let $f(x) = c_0x^3 + c_1x^2 + c_2x + c_3 (c_0 \neq 0)$ be an arbitrary cubic polynomial with real coefficients. Let $T_1(x, a) = f(a) + f'(a)(x - a)$ be the first degree Taylor polynomial associated with $f(x)$, so that the tangent line to the graph of $f(x)$ at $x = a$ is given by $y = T_1(x, a)$. The tangency of this line to

$y = f(x)$ at $x = a$ is equivalent to the fact that the polynomial $f(x) - T_1(x, a)$ is divisible by $(x - a)^2$. There is a unique remaining zero b of this cubic polynomial, given by $b = -c_1/c_0 - 2a$ (since the sum of the zeros is given by $-c_1/c_0$), giving the x-coordinate of the remaining point of intersection of the tangent line with the graph of $f(x)$. These two points of intersection are distinct except in the case where the tangent line intersects in a contact point of degree 3, i.e., an inflection point. This occurs when $-c_1/c_0 - 2a = a$, i.e., when $a = -c_1/(3c_0)$. At all points on the graph that are not inflection points, these two points of intersection define a region $R(a)$ bounded by the cubic and the tangent line between these two points. Let $A(a)$ denote the area of the region $R(a)$. Iterating this process, we can then define the region $R(b)$ with area $A(b)$ (note that if the point with $x = a$ is not an inflection point then neither is the point with $x = b$).

Let $\rho(a) = A(b)/A(a)$ denote the ratio of the area of $R(b)$ to the area of $R(a)$.

It is geometrically clear that $\rho(a)$ is not changed if the graph of $f(x)$ is scaled, i.e., if $f(x)$ is replaced by $cf(x)$ or by $f(cx)$ for any nonzero real number c. Similarly, $\rho(a)$ is invariant under translations, i.e., by replacing $f(x)$ by $f(x - c)$ or by $f(x) - c$ for any real number c. Finally, since the function $f(x) - T_1(x, a)$ is not changed if $f(x)$ is replaced by $f(x) + sx + t$ for any linear function $sx + t$, $\rho(a)$ is also invariant under any such transformation on $f(x)$.

Theorem 1. *Let $f(x)$ be any cubic polynomial with real coefficients. Then $\rho(a) = 16$ for all a for which $(a, f(a))$ is not an inflection point on the graph $y = f(x)$.*

Proof: By the preceding remarks, we may first translate x so that the coefficient of x^2 in $f(x)$ is zero. By translating $f(x)$ by a linear function, we may assume that $f(x)$ has no linear terms. Finally, scaling the resulting function so that its leading coefficient is 1 reduces the problem to consideration of $f(x) = x^3$. In this case $f(x) - T_1(x, a) = x^3 - a^3 - 3a^2(x - a) = (x - a)^2(x + 2a)$, so $b = -2a$, as mentioned in the Introduction. The area $A(a)$ of the region $R(a)$ between $y = x^3$ and $y = a^3 + 3a^2(x - a)$ from $x = a$ to $x = -2a$ is given by the absolute value of $\int_a^{-2a}(x - a)^2(x + 2a)\,dx = -27a^4/4$. Then $A(b) = A(-2a)$ is given by $27(-2a)^4/4$, so the ratio of these areas is $\rho(a) = 16$. □

3. GENERALIZATION TO POLYNOMIALS OF ARBITRARY DEGREE. The first difficulty encountered in trying to generalize the results of the previous section to other functions, in particular to polynomials $f(x) = c_0x^n + c_1x^{n-1} + \cdots + c_n$ of arbitrary degree n, is that in general the tangent line to $f(x)$ at a given point does not intersect the curve $y = f(x)$ in another uniquely defined point. One possible extension that overcomes this difficulty is to replace the tangent line with the graph of a higher order Taylor polynomial approximation to $f(x)$. Let $T_{n-2}(x, a)$ denote the Taylor polynomial approximation of degree $n - 2$ to $f(x)$ at $x = a$, the graph of which in the case $n = 3$ of cubics is just the tangent line to $f(x)$ at $x = a$ considered in the previous section.

The values of $f(x)$ and $T_{n-2}(x, a)$ together with their first $n - 2$ derivatives agree at $x = a$, so that

$$f(x) - T_{n-2}(x, a) = c_0(x - a)^{n-1}(x - b)$$

for some b, i.e., there is a unique additional point of intersection of $y = f(x)$ and $y = T_{n-2}(x, a)$. Computing the $(n - 1)$-st derivative at a shows that $f^{(n-1)}(a) = c_0(a - b)(n - 1)!$, and so $b = -c_1/c_0 - (n - 1)a$.

If $b \neq a$, i.e., if $a \neq -c_1/(nc_0)$ (so $b \neq -c_1/(nc_0)$), then the graphs of $f(x)$ and $T_{n-2}(x, a)$ between $x = a$ and $x = b$ define a finite region $R(a)$ in the plane. The area $A(a)$ of this region is given by the absolute value of $\int_a^b (x-a)^{n-1}(x-b)\,dx$. Integration by parts shows that $A(a) = -(b-a)^{n+1}/(n(n+1)) = -(-c_1/c_0 - na)^{n+1}/(n(n+1))$. The area $A(b)$ of the region $R(b)$ is then given by the absolute value of the same expression with a replaced by b. Since $b = -c_1/c_0 - (n-1)a$, it is easy to check that $c_1/c_0 + nb = -(n-1)(c_1/c_0 + na)$ and so the ratio of these two areas is $\rho(a) = A(b)/A(a) = (n-1)^{n+1}$. We summarize this as

Theorem 2. *Let $f(x)$ be a polynomial of degree n and let $T_{n-2}(x, a)$ be the Taylor polynomial approximation of degree $n-2$ to $f(x)$ at $x = a$. Then successive areas between $f(x)$ and $T_{n-2}(x, a)$ are proportional, with proportionality constant $\rho = \rho(a) = (n-1)^{n+1}$ depending only on the degree of $f(x)$.*

When $n = 3$, we have $\rho(a) = 2^4 = 16$, agreeing with Theorem 1.

4. THE ODD EXTENSION OF THE FUNCTION x^α. Another possible extension of the results for cubic polynomials (which ultimately reduces to the case $f(x) = x^3$) is to consider the graphs of the functions x^α for real $\alpha > 0$. This function is in general defined only for $x > 0$, but we can extend x^α to a continuous function defined for all real x by defining

$$F_\alpha(x) = \begin{cases} x^\alpha, & \text{for } x \geq 0 \\ -(-x)^\alpha, & \text{for } x < 0, \end{cases} \tag{1}$$

i.e., $F_\alpha(x)$ is the extension of x^α for $x \geq 0$ to an odd function of x. If α is the quotient of two odd integers, then $F_\alpha(x)$ is just x^α.

We shall first see that, for all $\alpha \neq 1$, the graph $y = F_\alpha(x)$ again has the property that tangent lines at all points $a \neq 0$ intersect the curve in precisely one additional point:

Proposition 1. *Let $\alpha \neq 1$ be a positive real number and let $F_\alpha(x)$ be defined by equation (1). Then $P(x) = x^\alpha - \alpha x - \alpha + 1$ has a unique positive zero $x = \omega(\alpha)$ and this zero satisfies $\omega(\alpha) > 1$. For all $a \neq 0$ the tangent line to the curve $y = F_\alpha(x)$ at $x = a$ intersects the curve in precisely one additional point, namely the point with $x = -\omega(\alpha)a$.*

Proof: Suppose first that $\alpha > 1$. Then $P(1) = 2 - 2\alpha < 0$ and $P(x)$ tends to $+\infty$ as x tends to $+\infty$, so there is a zero of $P(x)$ in the interval $(1, \infty)$. Then $P'(x) = \alpha(x^{\alpha-1} - 1)$ is zero only for $x = 1$, so there are no additional zeros for $x > 1$. For $0 < x < 1$, $P'(x)$ is negative, so $P(x)$ is decreasing in this interval. Since $\lim_{x \to 0^+} P(x) = 1 - \alpha < 0$, it follows that $P(x)$ is strictly negative for $0 < x \leq 1$, and in particular $P(x)$ has no additional zeros for $x > 0$.

The proof for $0 < \alpha < 1$ is similar. In this case $P(1) > 0$ and $P(x)$ tends to $-\infty$ as x tends to $+\infty$, so again there is a zero of $P(x)$ in the interval $(1, \infty)$. In this case $P'(x) > 0$ for $0 < x < 1$, and $\lim_{x \to 0^+} P(x) = 1 - \alpha > 0$, so $P(x)$ is strictly positive for $0 < x \leq 1$ and again there are no additional zeros for $x > 0$.

To prove the second statement in the proposition, suppose first that $a > 0$. Then the tangent line to the curve $y = F_\alpha(x)$ at $x = a$ has equation $y = T_1(x, a) = a^\alpha + \alpha a^{\alpha-1}(x - a)$. This tangent line does not intersect $y = F_\alpha(x)$ for

any additional $x \geq 0$ (since $y = F_\alpha(x)$ is concave up for $\alpha > 1$ and concave down for $0 < \alpha < 1$). The tangent line intersects $y = -(-x)^\alpha$ for $x < 0$ when $-(-x)^\alpha = a^\alpha + \alpha a^{\alpha-1}(x-a)$. This is equivalent to $(-x/a)^\alpha - \alpha(-x/a) - \alpha + 1 = 0$, i.e., $P(-x/a) = 0$, which occurs precisely for $-x/a = \omega(a)$, proving the proposition for $a > 0$. The case when $a < 0$ follows by symmetry. □

As before, let $R(a)$ be the region bounded by the graph of $F_\alpha(x)$ and its tangent line at $x = a$ between $x = a$ and $x = b = -\omega(\alpha)a$, let $A(a)$ denote its area, and let $\rho(a)$ denote the ratio $A(b)/A(a)$ of the successive areas. The next result proves that for all $\alpha \neq 1$ the ratio $\rho(a)$ for the curve $y = F_\alpha(x)$ has properties similar to those of the cubic case $\alpha = 3$.

Theorem 3. *Let $\alpha \neq 1$ be a positive real number, let $F_\alpha(x)$ be the odd extension of the function x^α defined in equation* (1), *and let $\omega(\alpha)$ be as in Proposition* 1. *Then, for $a \neq 0$, successive tangent lines to $y = F_\alpha(x)$ at $x = a$ enclose proportional areas, with proportionality constant $\rho = \rho(a) = \omega(\alpha)^{\alpha+1}$, independent of a.*

Proof: Note first that by symmetry, $A(a) = A(|a|)$ for all a. If $a > 0$ then the area $A(a)$ is given by the absolute value of

$$\int_{-\omega(\alpha)a}^{0} \left[-(-x)^\alpha - T_1(x,a)\right] dx + \int_0^a \left[x^\alpha - T_1(x,a)\right] dx$$

with $T_1(x,a) = \alpha a^{\alpha-1}x + (\alpha-1)a^\alpha$. Making the change of variables from x to $-ax$ in the first integral and from x to ax in the second shows that

$$A(a) = a^{\alpha+1}\left[\int_0^{\omega(\alpha)} (-x^\alpha + \alpha x + \alpha - 1)\, dx + \int_0^1 (x^\alpha - \alpha x + \alpha - 1)\, dx\right],$$

where the factor multiplying $a^{\alpha+1}$ is independent of a. It follows that for all $a \neq 0$, the quotient $A(-\omega(\alpha)a)/A(a)$ is given by $|-\omega(\alpha)a|^{\alpha+1}/|a|^{\alpha+1} = \omega(\alpha)^{\alpha+1}$. □

Note that when $\alpha = 3$, the function $F_\alpha(x)$ is just x^3. In this case the function $P(x)$ is the polynomial $x^3 - 3x - 2$, whose zeros are $-1, -1, 2$, so that $\omega(3) = 2$ and $\rho(3) = 2^4$, as determined in Section 2.

5. BEHAVIOR OF THE PROPORTIONALITY CONSTANT ρ AS A FUNCTION OF α. Recall from the previous section that for all positive $\alpha \neq 1$ the value $\omega(\alpha)$ is defined to be the unique positive zero of the function $P(x) = P(x, \alpha)$ defined in Proposition 1 and that the proportionality constant $\rho = \rho(\alpha)$ is given by $\rho(\alpha) = \omega(\alpha)^{\alpha+1}$. The function $\omega(\alpha)$ is a continuous function of α in the intervals $(0,1)$ and $(1,\infty)$. The next result shows that $\omega(\alpha)$ can be extended to a continuous function for all $\alpha > 0$ and indicates some properties of $\omega(\alpha)$.

Theorem 4. *Let β denote the unique real root of the equation $\beta \ln \beta = \beta + 1$ (so $\beta \approx 3.59112147667\ldots$) and let $\omega(\alpha)$ be as in Proposition* 1. *Then*

a. $\lim_{\alpha \to 1} \omega(\alpha) = \beta$, *so defining $\omega(1) = \beta$ extends $\omega(\alpha)$ to a continuous function defined for all $\alpha > 0$.*
b. $\omega(1/\alpha) = \omega(\alpha)^\alpha$ *for all $\alpha > 0$.*
c. *$\omega(\alpha)$ is a monotonically decreasing function of α.*
d. $\lim_{\alpha \to 0^+} \omega(\alpha) = \infty$ (*in fact, $\omega(\alpha) \sim 2/\alpha$ as α tends to* 0) *and* $\lim_{\alpha \to \infty} \omega(\alpha) = 1$.

Proof: The number $\omega(\alpha)$ satisfies the equation

$$\omega(\alpha)^{\alpha} - \alpha\omega(\alpha) - \alpha + 1 = 0$$

for all $\alpha > 0, \alpha \neq 1$. Replacing α by $1/\alpha$ and multiplying by α results in the equation $(\omega(1/\alpha)^{1/\alpha})^{\alpha} - \alpha\omega(1/\alpha)^{1/\alpha} - \alpha + 1 = 0$. Hence $\omega(1/\alpha)^{1/\alpha}$ is also a zero of $P(x)$. The uniqueness in Proposition 1 then forces $\omega(1/\alpha)^{1/\alpha} = \omega(\alpha)$, which proves (b).

Since $\beta^{\alpha} = \beta e^{(\alpha-1)\ln\beta}$, the function $P(\beta, \alpha) = \beta^{\alpha} - \alpha\beta - \alpha + 1$ can be expanded in a Taylor series valid for all $\alpha > 0$:

$$\begin{aligned} P(\beta, \alpha) &= \beta \sum_{i=0}^{\infty} \frac{(\ln \beta)^i}{i!}(\alpha - 1)^i - \alpha\beta - \alpha + 1 \\ &= (\beta \ln \beta - \beta - 1)(\alpha - 1) + \beta \sum_{i=2}^{\infty} \frac{(\ln \beta)^i}{i!}(\alpha - 1)^i \\ &= \frac{1}{2}\beta(\ln \beta)^2(\alpha - 1)^2 + \cdots, \end{aligned}$$

since $\beta \ln \beta = \beta + 1$. It follows that $P(\beta, \alpha) > 0$, since all the terms in this series are positive. Since $P(1, \alpha) < 0$, the root $\omega(\alpha)$ must lie in the interval $(1, \beta)$ for $\alpha > 1$.

Similarly, if $1 < x_0 < \beta$ is any fixed value of x, we can consider the Taylor series expansion of $P(x_0, \alpha)$ valid for $\alpha > 0$ to see that $P(x_0, \alpha) = (x_0 \ln x_0 - x_0 - 1)(\alpha - 1) + O((\alpha - 1)^2)$. The value $x_0 \ln x_0 - x_0 - 1$ is negative since $x_0 < \beta$ (the derivative of the function $t \ln t - t - 1$ is $\ln t$ so the function is monotonically increasing and is zero when $t = \beta$ by definition of β). It follows that $P(x_0, \alpha)$ is negative for $\alpha > 1$ and sufficiently close to 1. In particular, given any $\varepsilon > 0$, there exists an $\alpha_0 > 1$ such that for any α in the interval $(1, \alpha_0)$ we have $P(\beta - \varepsilon, \alpha) < 0$. Hence for $\alpha \in (1, \alpha_0)$, the zero $\omega(\alpha)$ lies in the interval $(\beta - \varepsilon, \beta)$, so that $\lim_{\alpha \to 1^+} \omega(\alpha) = \beta$. By the functional equation proved in (b), it follows that $\lim_{\alpha \to 1^-} \omega(\alpha) = \lim_{\alpha \to 1^+} \omega(1/\alpha) = \lim_{\alpha \to 1^+} \omega(\alpha)^{\alpha} = \beta$. This proves (a).

For $\alpha > 0, \alpha \neq 1$, the number $\omega = \omega(\alpha)$ satisfies the equation

$$\omega^{\alpha} - \alpha\omega - \alpha + 1 = 0. \tag{2}$$

Differentiating implicitly gives

$$\frac{d\omega}{d\alpha} = \frac{1 + \omega - \omega^{\alpha} \ln \omega}{\alpha(\omega^{\alpha-1} - 1)}, \tag{3}$$

valid for all $\alpha > 0, \alpha \neq 1$. Let $\psi(\alpha) = \omega(1/\alpha) = \omega(\alpha)^{\alpha}$. From equation (2) we have $\psi = \alpha\omega + \alpha - 1$. Differentiating and using equation (3) gives

$$\frac{d\psi}{d\alpha} = \frac{\omega^{\alpha-1}(1 + \omega - \omega \ln \omega)}{\omega^{\alpha-1} - 1}. \tag{4}$$

As noted above, the function $1 + t - t \ln t$ is positive when $1 < t < \beta$ and negative for all $t > \beta$. Since we have proved $1 < \omega < \beta$ for $\alpha > 1$, it follows that $d\psi/d\alpha > 0$ for $\alpha > 1$. From $\psi(1) = \omega(1) = \beta$ we conclude that $\psi(\alpha) > \beta$ for all $\alpha > 1$, and so $\omega(\alpha) = \psi(1/\alpha) > \beta$ for $0 < \alpha < 1$. It now follows that both numerator and denominator in equation (4) are negative for $0 < \alpha < 1$, so $d\psi/d\alpha > 0$ also for $0 < \alpha < 1$. Hence ψ is a monotonically increasing function for all $\alpha > 0, \alpha \neq 1$. Therefore $\omega(\alpha) = \psi(1/\alpha)$ is monotonically decreasing for $\alpha \neq 1$

and consequently is decreasing for all $\alpha > 0$ since $\omega > \beta$ for $\alpha < 1$ and $\omega < \beta$ for $\alpha > 1$. This proves (c).

Consider now $P(1 + 2/\sqrt{n}, n) = (1 + 2/\sqrt{n})^n - n(1 + 2/\sqrt{n}) - n + 1$ for a positive integer n. Expanding by the binomial theorem shows that $P(1 + 2/\sqrt{n}, n) > 0$ for $n > 2$. Since $P(1, n) = 2 - 2n < 0$ for $n > 2$, the zero $\omega(n)$ must lie between 1 and $1 + 2/\sqrt{n}$. It follows that $\lim_{\alpha \to \infty} \omega(\alpha) = 1$, which proves the second statement in (d). Since $\omega(1/\alpha) = \alpha\omega + \alpha - 1$, it now follows that $\omega(\alpha) \sim 2/\alpha$ as α tends to 0. □

By the first part of Theorem 4, the proportionality constant $\rho(\alpha) = \omega(\alpha)^{\alpha+1}$ defined for positive $\alpha \neq 1$ in Theorem 3 can be extended to a continuous function defined for all real $\alpha > 0$. For a fixed α it is not difficult to solve equation (2) numerically for ω and then plot the corresponding proportionality constant $\rho(\alpha)$. This plot suggested the results in the following theorem.

Theorem 5. *As in Theorem* 4, *let* β *denote the unique real root of the equation* $\beta \ln \beta = \beta + 1$ *and let* $\rho(\alpha)$ *be the proportionality constant in Theorem* 3. *Then*

a. $\rho(\alpha) = \rho(1/\alpha)$.
b. ρ *is monotonically increasing for* $\alpha > 1$ *and monotonically decreasing for* $0 < \alpha < 1$, *with an absolute minimum at* $\alpha = 1$, *where* $\rho(1) = \beta^2 \approx 12.89615346019\ldots$.
c. $\lim_{\alpha \to 0^+} \rho(\alpha) = \lim_{\alpha \to \infty} \rho(\alpha) = \infty$ (*in fact,* $\rho(\alpha) \sim 2\alpha$ *as* α *tends to* ∞).

Proof: We have $\rho(\alpha) = \omega(\alpha)^{\alpha+1} = \omega(\alpha)\omega(\alpha)^{\alpha} = \omega(\alpha)\omega(1/\alpha)$ by (b) in Theorem 4. Then (a), (c), and the value of $\rho(1)$ in (b) follow immediately from the results in Theorem 4. By (a), it suffices to prove that $\rho(\alpha)$ is monotonically increasing for $\alpha > 1$. This is (perhaps surprisingly) rather more difficult and, at the suggestion of the referees, we only briefly indicate a proof.

The derivative $d\rho/d\alpha$ for $\alpha \neq 1$ can be computed from equations (2) and (3) and an application of L'Hôpital's rule shows that $d\rho/d\alpha$ is continuous also for $\alpha = 1$ and equals 0 at this point. The difficult point is to verify that $d\rho/d\alpha > 0$ for $\alpha > 1$. This can be seen by using equation (2) to find the degree 6 Taylor polynomial approximation for α as a function of ω near $\omega = \beta$ and then using this Taylor polynomial approximation to estimate $d\rho/d\alpha$. It is interesting to note that the lower order Taylor polynomial approximations are not sufficient for the estimates required; in fact, they provide inequalities *opposite* from those necessary to prove the theorem. □

Remark. The fact that $\rho(\alpha) = \rho(1/\alpha)$ can also be seen directly: the graph $y = F_\alpha(x)$ is the reflection in the line $y = x$ of the graph of $y = F_{1/\alpha}(x)$, so by symmetry α and $1/\alpha$ have the same proportionality constant.

As we have seen, successive tangents to the graphs of the functions $F_\alpha(x)$ extending the functions x^α for positive $\alpha \neq 1$ define regions with proportional areas depending only on α. By Theorem 5, as α tends to 1, this proportionality constant approaches the number β^2 and this is the absolute minimum proportionality constant for any of these functions. On the other hand, the graphs of the corresponding functions $F_\alpha(x)$ are (non-uniformly) approaching the graph $y = x$, for which one cannot even define the regions being considered. We shall see in the next section that there is a function having the successive tangent property with ratio β^2.

6. CONVERSE QUESTIONS. Following the exhortation of Richard Courant that "man muß immer umkehren" ("always consider the converse")[1] we consider the interesting question of determining *all* functions $f(x)$ whose successive tangent lines enclose proportional areas. Considering the converse will in particular lead us to a function whose proportionality constant is the limiting value β^2.

The functions $F_\alpha(x)$ considered in Section 4 also satisfy the additional condition that for all $a \neq 0$ the tangent line at $x = a$ intersects the curve $y = F_\alpha(x)$ at a point with x-coordinate a constant multiple of a. In this section we shall classify all reasonably continuous functions having both of these two properties. In particular we shall see that these properties essentially characterize the functions $F_\alpha(x)$ defined by equation (1).

We first impose a smoothness condition on $f(x)$:

(a) The third derivative of $f(x)$ exists for all $x \neq 0$. In particular, $f(x)$ is defined and continuous for all $x \neq 0$.

We next require the tangent line condition evidenced by all the functions $F_\alpha(x)$:

(b) There is a nonzero constant $\kappa \neq 1$ such that, for all $a \neq 0$, the tangent line to $y = f(x)$ at $x = a$ exists and intersects $y = f(x)$ at a finite number of points, one of which has x-coordinate κa.

For $f(x)$ satisfying (a) and (b), let $G(a) = \int_a^{\kappa a}[f(x) - T_1(x, a)]\,dx$ be the integral giving the (signed) area from a to κa of the region between $f(x)$ and its tangent line $y = T_1(x, a)$ at $x = a$ (interpreted as an improper integral if the integration includes the point $x = 0$). We last require these areas to be proportional for successive tangent lines:

(c) There is a nonzero constant σ such that $G(\kappa a) = \sigma G(a)$ for all $a \neq 0$.

It is easy to check that if $f(x)$ satisfies (a), (b), and (c) then so does any function obtained by adding a linear function to $f(x)$ or by scaling $f(x)$, i.e., a function of the form $cf(Cx) + ax + b$. The next theorem shows that if σ is positive then, up to such transformations, $f(x)$ is either the function $F_\alpha(x)$ for some $\alpha \neq 1$ or is the function $x \ln|x|$.

Theorem 6. *Let $f(x)$ be a function satisfying the preceding conditions (a), (b), and (c) and let $\alpha = (\sigma - \kappa)/(\kappa(\kappa - 1))$.*

(1) *If $\sigma > 0$, then $\alpha > 0$ and there exists a nonzero constant c and constants c_0, c_1 such that*
  - (a) *for $\alpha \neq 1$, $f(x) = cF_\alpha(x) + c_1 x + c_0$, where $F_\alpha(x)$ is the odd extension of the function x^α defined in equation (1) of Section 4, and*
  - (b) *for $\alpha = 1$, $f(x) = cx \ln|x| + c_1 x + c_0$.*

In both cases, $f(x)$ is defined and continuous for all real x, with $-\kappa = \omega(\alpha)$ and $\sigma = \rho(\alpha)$ for the functions ω and ρ of the previous section.

[1]This bit of mathematical wisdom was uttered by Courant at the blackboard after getting stuck on a proof while substituting for a sick Lipman Bers at a course attended by one of the authors at the NYU Institute of Mathematical Sciences (later to be renamed the Courant Institute). Considering the converse gave Courant the key to solving his problem that day; his statement remains as excellent advice.

(2) *If* $\sigma < 0$, *then* $-1 < \alpha < 1$ *and there exists a nonzero constant* c *and constants* c_0, c_1 *such that*

(c) *for* $\alpha \neq 0, f(x) = c|x|^\alpha + c_1 x + c_0$, *and*

(d) *for* $\alpha = 0, f(x) = c \ln|x| + c_1 x + c_0$.

In case (c), *the constant* $-\kappa$ *is the unique positive root of the equation* $x^\alpha + \alpha x + \alpha - 1 = 0$ *and satisfies* $0 < -\kappa < \sqrt{2} - 1$. *In this case,* $\sigma = -(-\kappa)^{\alpha+1}$ *and satisfies* $1 - \sqrt{2} < \sigma < 0$. *In case* (d), $-\kappa = \beta_0 \approx 0.27846\cdots$ *where* $1 + \beta_0 + \ln \beta_0 = 0$ *and* $\sigma = -\beta_0$.

Proof: For brevity we shall restrict attention to the case where $\sigma > 0$; the computation for $\sigma < 0$ is similar. For all $a \neq 0$, let $g(x, a) = f(x) - T_1(x, a) = f(x) - f(a) - f'(a)(x - a)$, so that $g(a, a) = g(\kappa a, a) = 0$ by assumption (b). Differentiating the relation $G(a) = \int_a^{\kappa a} g(x, a)\, dx$ with respect to a gives

$$G'(a) = \kappa g(\kappa a, a) - g(a, a) + \int_a^{\kappa a} \frac{\partial g(x, a)}{\partial a} dx$$

$$= \int_a^{\kappa a} \{-f'(a) - [f''(a)(x - a) - f'(a)]\}\, dx$$

$$= -\frac{1}{2} f''(a)(x - a)^2 \Big|_a^{\kappa a} = -\frac{1}{2}(\kappa - 1)^2 a^2 f''(a),$$

valid for all $a \neq 0$. Differentiating the relation in hypothesis (c) gives $\kappa G'(\kappa a) = \sigma G'(a)$, and using the previous result for a and for κa shows that

$$f''(\kappa a) = \frac{\sigma}{\kappa^3} f''(a) \tag{5}$$

for all $a \neq 0$. Now, the hypothesis in (b) that the point $(\kappa a, f(\kappa a))$ lies on the tangent line $y = T_1(x, a)$ is equivalent to

$$f(\kappa a) = f(a) + (\kappa - 1) a f'(a). \tag{6}$$

Differentiating this equation twice with respect to a gives

$$\kappa^2 f''(\kappa a) = (2\kappa - 1) f''(a) + (\kappa - 1) a f^{(3)}(a).$$

Then equation (5) implies that, for all $a \neq 0$,

$$a f^{(3)}(a) = (\alpha - 2) f''(a) \tag{7}$$

where α is the constant $(\sigma - \kappa)/(\kappa(\kappa - 1))$ of the theorem.

Let $F(x) = |x|^\alpha$, defined for all $x \neq 0$ and all α. Then $aF^{(3)}(a) = (\alpha - 2)F''(a)$ for all $a \neq 0$. If $\alpha \neq 0, 1$ then $F''(a)$ is never 0 and it follows from equation (7) that $d(f''/F'')/da = 0$ for all $a \neq 0$. Hence $f''(a)$ is a constant multiple of $F''(a)$ for all $a > 0$ and also for all $a < 0$, possibly with different constants. It follows, for $\alpha \neq 0, 1$, that there are constants c, c', c_0, c_0', c_1, and c_1' such that

$$f(x) = \begin{cases} c|x|^\alpha + c_1 x + c_0 & \text{for } x > 0 \\ c'|x|^\alpha + c_1' x + c_0' & \text{for } x < 0 \end{cases}. \tag{8}$$

The constants c and c' are nonzero by the finiteness condition in hypothesis (b). The function $|x|^\alpha$ is concave for $x > 0$, so it follows from hypothesis (b) that $\kappa < 0$. Then equation (6) for $a > 0$ becomes

$$c'(-\kappa)^\alpha a^\alpha + c_1' \kappa a + c_0' = (c a^\alpha + c_1 a + c_0) + (\kappa - 1) a (\alpha c a^{\alpha - 1} + c_1).$$

For $\alpha \neq 0, 1$, the functions 1, a, and a^α are linearly independent, so comparing

coefficients of these functions in the previous equation shows that $c'_0 = c_0$, $c'_1 = c_1$, and $c'|\kappa|^{\alpha} = c[1 + \alpha(\kappa - 1)] = c\sigma/\kappa$. Considering equation (6) for $a < 0$ shows that also $c|\kappa|^{\alpha} = c'\sigma/\kappa$, from which it follows that $(\sigma/\kappa)^2 = |\kappa|^{2\alpha}$, i.e., $\sigma = \pm|\kappa|^{\alpha+1}$.

If σ is positive, then $\sigma = |\kappa|^{\alpha+1}$ and then $c' = -c$. Further, since κ is negative, it follows that $\alpha = (\sigma - \kappa)/(\kappa(\kappa - 1))$ is positive. Hence $f(x) = cF_{\alpha}(x) + c_1x + c_0$ with $-\kappa = \omega(\alpha)$ and $\sigma = \rho(\alpha)$, as in the previous section.

For the case $\alpha = 1$, equation (7) becomes $d(af'')/da = 0$, so that $f(x) = cx \ln x + c_1x + c_0$ for $x > 0$ and $f(x) = c'x \ln(-x) + c'_1x + c'_0$ for $x < 0$ for some nonzero constants c, c' and constants c_0, c_1, c'_0, c'_1. Concavity again shows that $\kappa < 0$, and equation (6) for $a < 0$ together with the linear independence of the functions 1, a, and $a \ln a$ shows that $c' = c$, $c'_0 = c_0$, and $c[\kappa \ln(-\kappa) - \kappa + 1] = \kappa(c_1 - c'_1)$. Equation (6) for $a < 0$ shows that the latter equation also holds with c_1 and c'_1 interchanged. Both sides of the equation must therefore be 0, which proves that $c'_1 = c_1$ and that $-\kappa = \beta = \omega(1)$. Hence $f(x) = cx \ln|x| + c_1x + c_0$. Since $\alpha = 1$, we have $\sigma = \kappa^2 = \beta^2 = \rho(1)$.

For the case $\alpha = 0$ a similar argument shows that $\sigma < 0$ and leads to the functions in (d) of the theorem.

Remark. For the functions in part (1) of Theorem 6, the regions defined by successive tangents increase in size, and for the functions in part (2) they decrease.

Theorem 6 shows that (with the additional hypothesis (b)) the odd extensions of the functions x^{α} with $\alpha \neq 1$ (or $\sigma \neq \kappa^2$) considered in Sections 4 and 5 are essentially the only reasonably continuous functions whose successive tangents define regions with proportional (signed) areas with a positive proportionality constant. Furthermore, the proof of Theorem 6 shows that the odd extension of the function $x \ln x$ is the natural solution to this problem with $\alpha = 1$ ($\sigma = \kappa^2$) and by the results of the previous section, its proportionality constant is the limit of the proportionality constants for the functions with $\alpha \neq 1$. In this sense, the limit of the functions x^{α} as α tends to 1 is $x \ln x$ and not the function x!

It would be of interest to determine whether there are functions whose successive tangent lines enclose proportional areas other than those considered here, for example functions that do not satisfy the smoothness condition (a) or the tangent line condition (b).

ACKNOWLEDGMENTS. We would like to acknowledge some useful conversations with Richard Foote and Jack Lawlor.

Mathematics Department, ES-127A
The University of Albany
Albany, NY 12222
hibrown@math.albany.edu

Department of Mathematics & Statistics
University of Vermont
16 Colchester Avenue
Burlington, VT 05401
burgmeie@math.uvm.edu
dummit@math.uvm.edu

On the Discriminant Criterion and a Generalization

M. H. Eggar

Given a positive integer k and a polynomial with complex coefficients

$$P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0, \tag{1}$$

what condition on $(a_0, \ldots, a_{n-1})$ is necessary and sufficient for the equation $P(z) = 0$ to have a root of multiplicity at least k? For the case $k = 2$, two classical approaches to multiple roots are discussed in Sections 1 and 2. These are reconciled and for general k the answer is described in Section 5.

1. THE DISCRIMINANT. The discriminant criterion $\Delta(a_0, \ldots, a_{n-1}) = 0$ answers the question for $k = 2$. The *discriminant* Δ of P is the product of the $\binom{n}{2}$ squared differences of the roots $\alpha_1, \ldots, \alpha_n$

$$\Delta = \prod_{i<j} (\alpha_i - \alpha_j)^2,$$

or, equivalently, the square of their Vandermonde determinant. The discriminant can be written as a polynomial $\Delta(a_0, \ldots, a_{n-1})$ in $a_0, \ldots, a_{n-1}$, since the defining product is a symmetric polynomial of the roots α_i and the main result on symmetric polynomials gives an algorithmic procedure for writing any symmetric polynomial in the α_i as a polynomial in the elementary symmetric polynomials $(-1)^i a_{n-i}$ of the α_i. Clearly Δ vanishes if and only if at least two roots coincide. Equivalently, from the Vandermonde point of view, the non-vanishing or otherwise of the determinant is the criterion for whether or not the n simultaneous linear equations in $a_0, \ldots, a_{n-1}$, $\alpha_i^n + \Sigma_r a_{n-r}\alpha_i^{n-r} = 0$ for $1 \leq i \leq n$, have a unique solution. For α_i not all distinct there is duplication amongst these equations and so the a_j are not uniquely determined, but for α_i all distinct the polynomial P is $\Pi_i(z - \alpha_i)$ and so the a_j are uniquely determined.

We note that if $P(z)$ is replaced by the general polynomial $a_n z^n + a_{n-1}z^{n-1} + \cdots + a_0$ of degree not exceeding n, then the condition for a root of multiplicity at least 2 splits into cases according to the precise degree. The condition is $\Delta(a_0/a_n, \ldots, a_{n-1}/a_n) = 0$ if $a_n \neq 0$, $\Delta(a_0/a_{n-1}, \ldots, a_{n-2}/a_{n-1}) = 0$ if $a_n = 0$, $a_{n-1} \neq 0$, etc. The whole condition cannot be described by the vanishing of a single expression in $a_0, \ldots, a_n$.

Returning to the polynomial $P(z)$ of (1) one can with more effort count the number, N say, of pairs (i, j) such that $i < j$ and $\alpha_i = \alpha_j$. One observes that z^N is the highest power of z that divides the polynomial $p(z)$ of degree $\binom{n}{2}$ with zeros $(\alpha_1 - \alpha_2)^2, (\alpha_1 - \alpha_3)^2, \ldots, (\alpha_{n-1} - \alpha_n)^2$. Each coefficient in $p(z)$ can be expressed in terms of $a_0, \ldots, a_{n-1}$, since the coefficients are with alternating signs

the elementary symmetric polynomials in the $\binom{n}{2}$ roots and hence are symmetric in $\alpha_1, \ldots, \alpha_n$. It is worth mentioning that the computation is greatly facilitated by a trick [1, p. 172] based on the observation that if λ is added to each α_i, $p(z)$ is unchanged whereas the a_i do change. The discriminant criterion is the statement that the constant term of $p(z)$ vanishes.

If one knows that all roots α_i are real it is easy to extend the discriminant criterion to detect the existence of a root of multiplicity at least k. For example, for $k = 3$ the product

$$\prod_{h<i<j} \left[(\alpha_h - \alpha_i)^2 + (\alpha_i - \alpha_j)^2 + (\alpha_j - \alpha_h)^2\right]$$

is symmetric in $\alpha_1, \ldots, \alpha_n$ and hence can be expressed as a polynomial in $a_0, \ldots, a_{n-1}$. The vanishing of this polynomial is equivalent to the existence of a root of multiplicity at least 3. Of course, this method fails if the roots need not be real, since a sum of squares of complex numbers may vanish without each square being zero. In the general case (of complex roots) the criterion on $(a_0, \ldots, a_{n-1})$ that detects existence of a root of multiplicity at least k cannot be described by a single equation and our task is to elucidate what form it does take. To do this we change our approach.

2. MULTIPLE ROOTS AND THE EUCLIDEAN ALGORITHM. A necessary and sufficient condition for α to be a root of multiplicity precisely k is that $P(\alpha) = P^1(\alpha) = P^2(\alpha) = \cdots = P^{k-1}(\alpha) = 0$ and $P^k(\alpha) \neq 0$, where $P^i(z)$ denotes the polynomial obtained by differentiating $P(z)$ i times with respect to z. Set $R_1(z) = P(z)$, $R_2(z) = \gcd(P(z), P^1(z))$, $R_3(z) = \gcd(R_2(z), P^2(z))$ and inductively $R_i(z) = \gcd(R_{i-1}(z), P^{i-1}(z))$, where gcd denotes greatest common divisor and may be constructed by the Euclidean algorithm. In other words, $R_i(z) = \gcd(P(z), P^1(z), \ldots, P^{i-1}(z))$ and so $R_i(z) = 1$ for all sufficiently large i. If $i \leq r, (z - \alpha)^r$ is a factor of $P(z)$ if and only if $(z - \alpha)^{r-i+1}$ is a factor of $R_i(z)$. Set $Q_i(z) = R_i(z) R_{i+2}(z)(R_{i+1}(z))^{-2}$; then no $Q_i(z)$ has a repeated factor and $(z - \alpha)$ is a factor of $Q_i(z)$ if and only if α is a root of $P(z) = 0$ with multiplicity precisely i. Thus from the Euclidean algorithm one obtains (constructively!) the factorization [1, p. 130]

$$P(z) = Q_1(z)(Q_2(z))^2(Q_3(z))^3 \cdots,$$

which separates the factors with different multiplicity and counts their numbers. Although this technique yields complete information for any given polynomial P, it does not directly answer our initial question.

3. EXAMPLES OF THE DISCRIMINANT OBTAINED VIA THE EUCLIDEAN ALGORITHMS. By the Euclidean algorithm approach in Section 2 the condition for existence of a repeated root is that $R_2(z)$ is not the constant polynomial. In the two simplest cases one easily recovers the discriminant condition.

Example 3.1. Suppose $P(z) = z^2 + az + b$. The Euclidean algorithm to calculate $\gcd(P(z), P^1(z))$ is

$$z^2 + az + b = (2z + a)\left(\frac{1}{2}z + \frac{1}{4}a\right) + \left(b - \frac{1}{4}a^2\right).$$

Hence $P(z)$ has a repeated root if and only if $a^2 - 4b = 0$.

Example 3.2. Suppose $P(z) = z^3 + az + b, a \neq 0$. The Euclidean algorithm to calculate $\gcd(P(z), P^1(z))$ is

$$z^3 + az + b = (3z^2 + a)\left(\frac{1}{3}z\right) + \left(\frac{2}{3}az + b\right)$$

$$3z^2 + a = \left(\frac{2}{3}az + b\right)\left(\frac{9}{2a}z - \frac{27b}{4a^2}\right) + \frac{27b^2 + 4a^3}{4a^2}.$$

Hence $P(z)$ has a repeated root if and only if $27b^2 + 4a^3 = 0$. The condition for a root of multiplicity at least 3 is that $\frac{2}{3}az + b$ is the zero polynomial and that $3z^2 + a$ has vanishing discriminant, i.e., $a = 0, b = 0$.

In more complicated examples, different cases can occur in the shape of the Euclidean algorithm that produces $R_2(z)$. For example, for $P(z) = z^5 + a_4z^4 + \cdots + a_0$ the first line of the algorithm reads

$$P(z) = P^1(z)\left(\frac{1}{5}z + \frac{1}{25}a_4\right) + r(z)$$

where $r(z) = \frac{1}{25}((10a_3 - 4a_4^2)z^3 + (15a_2 - 3a_3a_4)z^2 + (20a_1 - 2a_2a_4)z + (25a_0 - a_1a_4))$. The subsequent lines of the algorithm take a different form depending on, for example, whether or not $(10a_3 - 4a_4^2)$ vanishes. In general, for $P(z)$ of degree n one gets a tree of different cases depending on the degrees of the remainders in each line of the algorithm. There are 2^{n-1} cases, since the degree of the remainders can be any subsequence (see Lemma 4.2) of the descending sequence $\{n-2, n-3, \ldots, 0\}$. Here we interpret the null subsequence as the case $R_2(z) = P^1(z)$.

The methods of Section 2 thus split the criterion for the existence of a repeated root into 2^{n-1} cases. In Section 5 we will reconcile this with our knowledge from Section 1 that the criterion is the vanishing of a single expression, the discriminant.

4. THE EUCLIDEAN ALGORITHM FOR $\gcd(P, P^1)$. Let $d_1, d_2, \ldots, d_m$ be non-negative integers and let $P(z), Q(z)$ be polynomials of degree n, n' respectively, where $n \geq n' \geq m + \Sigma d_i$. We say that the Euclidean algorithm for P, Q has *deficiency* $(d_1, \ldots, d_m)$ if it takes the form

$$\begin{aligned}
P(z) &= Q(z)q_1(z) + r_1(z)\\
Q(z) &= r_1(z)q_2(z) + r_2(z)\\
r_1(z) &= r_2(z)q_3(z) + r_3(z)\\
&\ \vdots\\
r_{m-2}(z) &= r_{m-1}(z)q_m(z) + r_m(z)\\
r_{m-1}(z) &= r_m(z)q_{m+1}(z)
\end{aligned} \tag{2}$$

where $\deg r_1 = n' - 1 - d_1$ and $\deg r_i = \deg r_{i-1} - 1 - d_i$ for $2 \leq i \leq m$. Here $\deg r_i$ denotes the degree of the polynomial $r_i(z)$. In particular, $n' = c + m + \Sigma d_i$, where $c = \deg r_m$.

Lemma 4.2. *Given any non-negative integers $n, d_1, \ldots, d_m$ such that $n \geq 1 + m + \Sigma d_i$, there exist polynomials P of degree n such that the Euclidean algorithm for P, P^1 has deficiency $(d_1, \ldots, d_m)$.*

Proof: The key to the proof is to concentrate in (2) on the q_i rather than the r_i. Note that the conditions on $\deg r_i$ are equivalent to $\deg q_{i+1} = 1 + d_i$ for $1 \leq i \leq m$. Eliminating the r_i from the equations in (2), one obtains the continued

fraction expansion $[q_1, q_2, \ldots, q_{m+1}]$ for $\frac{P}{Q}$. Define $A_1 = q_1$, $B_1 = 1$, $A_2 = q_1q_2 + 1$, $B_2 = q_2$ and in general $A_r = q_rA_{r-1} + A_{r-2}$, $B_r = q_rB_{r-1} + B_{r-2}$. Standard results in the theory of continued fractions [2, p. 132] ensure that $\gcd(A_r, B_r) = 1$ and $P = r_mA_{m+1}, Q = r_mB_{m+1}$. Thus

$$P = q_1q_2 \cdots q_{m+1}r_m + \text{lower terms} \tag{3}$$

$$Q = q_2q_3 \cdots q_{m+1}r_m + \text{lower terms} \tag{4}$$

where the lower terms consist of a sum of certain monomials in the q_i and r_m. For P each monomial is a product of fewer than $m + 2$ of $q_1, \ldots, q_{m+1}, r_m$ and for Q is a product of fewer than $m + 1$ of $q_2, \ldots, q_{m+1}, r_m$.

Now take $n' = n - 1$ in (2) and let $q_1, \ldots, q_{m+1}, r_m$ be polynomials of degree $1, 1 + d_1, \ldots, 1 + d_m, c$ with general coefficients, where $c = n - 1 - m - \Sigma d_i$. For any polynomial $f(z)$ write $(f)_j$ for the coefficient of z^j. For example $q_1 = (q_1)_1z + (q_1)_0$. Imposition of the condition $Q = P^1$ imposes necessary and sufficient conditions on the coefficients as follows. Comparing the coefficient of z^{n-1} on both sides of $Q = P^1$, we get

$$\begin{aligned}(q_2q_3 \cdots q_{m+1}r_m)_{n-1} &= n(q_1 \cdots q_{m+1}r_m)_n \\ &= n(q_1)_1(q_2 \cdots q_{m+1}r_m)_{n-1},\end{aligned}$$

so $(q_1)_1 = \frac{1}{n}$. Choose any ordering of the $2m + \Sigma d_i$ coefficients $(q_i)_j$, where $2 \le i \le m + 1$, $0 \le j \le 1 + d_i$, such that j is non-increasing in the ordering. Continue this ordered list with $(q_1)_0$ and then with $(r_m)_0, (r_m)_1, \ldots, (r_m)_c$ in any order if $c \le 1 + m$ and in the order $(r_m)_{c-m-2}, (r_m)_{c-m-3}, \ldots, (r_m)_0, (r_m)_{c-m-1}, (r_m)_{c-m}, \ldots, (r_m)_c$ if $c > 1 + m$. From (3) and (4) we see that comparison of the coefficients of $z^{n-2}, \ldots, 1$ in $Q = P^1$ can be regarded as defining the first $n - 1 (= m + \Sigma d_i + c)$ elements of the ordered list in terms of elements that occur later in the list, indeed in terms of the last $(m + 2)$ elements, since $m + 2 = (2m + \Sigma d_i + c + 2) - (m + \Sigma d_i + c)$. If P is required to be monic we choose $(r_m)_c$ appropriately. ∎

Corollary 4.5. *The space of monic polynomials P of degree n such that the Euclidean algorithm for P, P^1 has deficiency $(d_1, \ldots, d_m)$ has $(m + 1)$ independent parameters. In particular, the number of parameters does not depend upon* $\deg(\gcd(P, P^1)$, *i.e., upon c.*

Corollary 4.6. *Suppose one formally writes out the Euclidean algorithm for P, P^1, where $P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$, successively imposing the conditions that the remainders r_i at the ith step have degrees given by $\deg r_1 = \deg P^1 - d_1 - 1$, $\deg r_{i+1} = \deg r_i - 1 - d_i$ for $i \ge 1$, and that r_{m+1} vanishes. That is, one sets the coefficients of the d_i potentially highest powers of z in r_i and all coefficients of r_{m+1} equal to zero. Then these conditions are algebraically independent.*

Proof: The number of closed conditions is $\Sigma d_i + c$, where $c = \deg r_m$. The conditions are algebraically independent if and only if the space of polynomials P satisfying them is determined by $n - (\Sigma d_i + c)$ independent parameters. By the construction in the proof of Lemma 4.2 this is the case, since $n - (\Sigma d_i + c) = m + 1$. ∎

Lemma 4.7. *Let S_c denote the space of monic polynomials P such that* $\gcd(P, P^1)$ *has degree c. Then there exists a subset T_c of S_c determined by $n - c$ independent parameters such that S_c is the closure of T_c.*

Proof: Take T_c to be the set of polynomials with $(n - c)$ distinct zeros, c of which are double zeros. The distinct roots can be regarded as the independent parameters. ■

Corollary 4.8. *Let $S_{c,0}$ denote the space of monic polynomials P such that* $\gcd(P, P^1)$ *has degree c and the Euclidean algorithm for P, P^1 has deficiency* $(0, 0, \ldots, 0)$ (*and hence $m = n - 1 - c$*). *Then every polynomial of S_c is in the closure of $S_{c,0}$.*

Proof: S_c is partitioned as a union of $S_{c,d}$, where $S_{c,d}$ denotes the subset of those polynomials with deficiency d. We remark that each $S_{c,d}$ is determined by some closed conditions and some inequalities. The inequalities assert that the coefficient of a certain power of z in $r_i(z)$ does not vanish and the closed conditions assert that the coefficients of potentially higher powers of z do vanish. By Corollary 4.5 only $S_{c,0}$ is given by $n - c$ independent parameters; the other $S_{c,d}$ have fewer parameters. Hence every polynomial P lies in the closure of $T_c \cap S_{c,0}$, where T_c is as in Lemma 4.7. ■

5. DISCRIMINANT CONDITIONS VIA THE EUCLIDEAN ALGORITHM. We are now in a position to describe a necessary and sufficient condition for the polynomial $P(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$ to have a root of multiplicity at least k. One writes out the Euclidean algorithm for P, P^1 assuming the remainder at each stage has maximal degree. This has the form (2) with $m = n - 1$ and $d_i = 0$ for all i. Each coefficient of each $r_i(z)$ will of course be an algebraic function of $a_0, \ldots, a_{n-1}$.

For $k = 2$ the condition is that the constant r_{n-1} vanishes. This is the discriminant condition.

For $k = 3$ the condition is that both coefficients of the linear function r_{n-2} vanish and that r_{n-3} is a perfect square.

For general k the condition is that all coefficients of r_{n-k-1} vanish and that r_{n-k-2} has the form $\lambda(z - \mu)^{k-1}$.

Note that these conditions are all given by the vanishing of certain functions of $a_0, \ldots, a_{n-1}$. Corollary 4.8 justifies the assertion that these closed conditions do cater for all polynomials, whether or not the Euclidean algorithm for P, P^1 has deficiency $(0, \ldots, 0)$. The condition must be closed since the roots are continuous functions of the coefficients a_i.

Finally we remark that the literature contains other generalizations of the discriminant. One of these is the condition for $P(z) = 0$ to have exactly k distinct roots, another is the "resultant" of two polynomials, which vanishes when the polynomials have a common zero.

REFERENCES

1. J. W. Archbold, *Algebra*, Pitman, 1964.
2. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Third Edition, Clarendon Press, Oxford, 1954.

Department of Mathematics and Statistics
The University of Edinburgh
Kings Building, Mayfield Road
Edinburgh EH9 3JZ, Scotland
m.eggar@ed.ac.uk

NOTES

Edited by: **John Duncan**

Expressing Primes as Quadratic Forms of Integers

Ernst Snapper

It is well known that if p is a rational prime and $p \equiv 1 \pmod 3$, there are rational integers a, b such that $p = a^2 + b^2 - ab$ [**1**, Proposition 8.3.1]. In this note, the prime 3 is replaced by an arbitrary odd prime q and a simple integral quadratic form with value p is derived when $p \equiv 1 \pmod q$.

1. THE QUADRATIC FORM WITH VALUE p. Everywhere, *integer* and *prime* stand for *rational integer* and *rational prime*.

Theorem. *If p and q are odd primes and $p \equiv 1 \pmod q$, there are $q-1$ integers $a_0, a_1, \ldots, a_{q-2}$ such that*

$$p = a_0^2 + a_1^2 + \cdots + a_{q-2}^2 - (a_0 a_1 + a_1 a_2 + \cdots + a_{q-3} a_{q-2}).$$

Proof: Since $p \equiv 1 \pmod q$, there is a character χ of order q on the field F_p with p elements. If $a \in F_p$, $\chi(a)$ is either 0 or a qth root of 1, and hence $\chi(a) \in \mathbb{Z}[\zeta]$ where ζ is a primitive qth root of 1. It follows that the Jacobi sum $J(\chi, \chi) = \sum_{a+b=1} \chi(a)\chi(b)$ is an element of $\mathbb{Z}[\zeta]$ and hence there are integers $a_0, a_1, \ldots, a_{q-2}$ such that

$$J(\chi, \chi) = a_0 + a_1 \zeta + \cdots + a_{q-2} \zeta^{q-2}.$$

The remainder of the proof shows that these integers have the required property.

Since both χ and χ^2 have order q and hence are not equal to the character of order 1, the absolute value of $J(\chi, \chi)$ is equal to $\sqrt{p}$ [**1**, p. 94, Corollary]. It follows that p is equal to the product of $J(\chi, \chi)$ and its complex conjugate and hence

$$p = \sum_{i=0}^{q-2} a_i \zeta^i \sum_{j=0}^{q-2} a_j \zeta^{-j} = \sum_{i,j=0}^{q-2} a_i a_j \zeta^{i-j}. \tag{1}$$

The last sum is also equal to $v_0 + v_1 \zeta + \cdots + v_{q-2} \zeta^{q-2}$ where $v_0, v_1, \ldots, v_{q-2}$ are integers, and since p is equal to this sum, $p = v_0$ and $v_1 = \cdots = v_{q-2} = 0$. Hence we need to show only that

$$v_0 = a_0^2 + a_1^2 + \cdots + a_{q-2}^2 - (a_0 a_1 + a_1 a_2 + \cdots + a_{q-3} a_{q-2}).$$

In the last sum of (1), put $i - j = u_{ij} q + r_{ij}$ where u_{ij} and r_{ij} are integers and $0 \le r_{ij} \le q-1$. Then this sum becomes $\sum_{i,j=0}^{q-2} a_i a_j \zeta^{r_{ij}}$ and the term $a_i a_j \zeta^{r_{ij}}$ contributes to v_0 only if $r_{ij} = 0$ or $q-1$. Clearly, $-q+2 \le i-j \le q-2$.

If $r_{ij} = 0, -q + 2 \le u_{ij}q \le q - 2$ and this can only happen if $u_{ij} = 0$. Then $i = j$ and the terms which contribute to v_0 form the sum $\sum_{i=0}^{q-2} a_i^2$.

If $r_{ij} = q - 1, -q + 2 \le (u_{ij} + 1)q - 1 \le q - 2$ and this can only happen if $u_{ij} = -1$. Then $i - j = -1$ and the terms that contribute to v_0 form the sum $\sum_{i=0}^{q-3} a_i a_{i+1} \zeta^{q-1}$. Since $\zeta^{q-1} = -(1 + \zeta + \cdots + \zeta^{q-2})$, the contribution of this sum to v_0 is $-\sum_{i=0}^{q-3} a_i a_{i+1}$ and hence $v_0 = \sum_{i=0}^{q-2} a_i^2 - \sum_{i=0}^{q-3} a_i a_{i+1}$. ∎

2. EXAMPLES. If $q = 3$, the theorem gives the classical result mentioned in the first paragraph.

If $q = 5$, $11 \equiv 1 \pmod 5$ and $11 = 2^2 + 3^2 + 4^2 - (2 \cdot 3 + 3 \cdot 4)$. Hence we can choose $a_0 = 2, a_1 = 3, a_2 = 4, a_3 = 0$.

Also, $31 \equiv 1 \pmod 5$ and $31 = 5^2 + 6^2 - (5 \cdot 6)$ whence we can choose $a_0 = 5, a_1 = 6, a_2 = a_3 = 0$.

REFERENCE

1. K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, second edition, Springer, New York, 1990.

Department of Mathematics
Dartmouth College
Hanover, New Hampshire 03755

Volumes of Cones

Richard J. Bagby

To find the volume of a cone, we don't necessarily need to find any of its dimensions. What we need is an appropriate product of them, and that can be significantly easier to find. Here's a beautiful application of this principle; it leads to a remarkable theorem about volumes of general truncated quadratic cones.

Start with a water-filled right circular cone, initially having its base on top. Call its radius R and altitude H. We now tip the cone and spill some of the water; we assume the remainder occupies an obliquely truncated cone with base B tangent to the rim. Our problem is to calculate its volume. Instead of tackling it directly, we build a new cone of equal volume. Calling L the axis of the original right circular cone, our new cone has its base A in a plane through the original vertex and perpendicular to L; we form A by projecting B. The new cone has its vertex where L meets B, so its altitude h is the distance between the two vertices. All three cones are sketched in Figure 1. We find it helpful to think of L as vertical rather than tipped, and have drawn it accordingly.

It's easy to see that we've preserved the volume. Calling θ the angle between L and the normal to B, the oblique cone has altitude $h \cos \theta$, and $\frac{1}{3} Bh \cos \theta = \frac{1}{3} Ah$ by the area cosine principle.

To find the area of A, first note that B has a plane of symmetry through L, and that plane contains the highest and lowest points of B. Calling r the distance to L

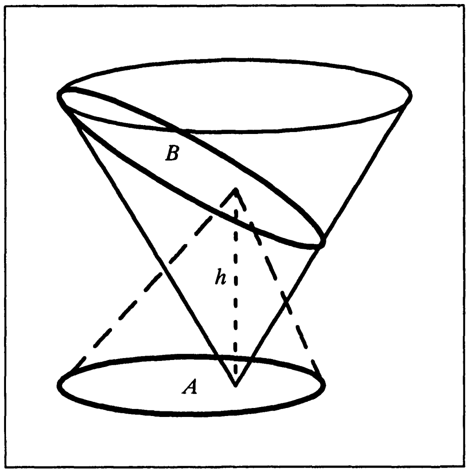


Figure 1. Cones of equal volume.

from the lowest point and defining $a = \frac{1}{2}(R + r)$, we choose coordinates for the plane of A to make the projections of these extreme points be $(-a, 0)$ and $(a, 0)$ respectively. Then L passes through $(c, 0)$, with $c = \frac{1}{2}(R - r)$. Since our projection preserves distances to L, the distance from $(c, 0)$ to a point (x, y) on the boundary of A is a linear function of x, namely

$$\rho(x) = \frac{R(a - x)}{2a} + \frac{r(a + x)}{2a} = a - \frac{c}{a}x,$$

so the boundary of A has the equation

$$(x - c)^2 + y^2 = \rho(x)^2,$$

and that simplifies readily to

$$\frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1.$$

Since $a^2 - c^2 = Rr$, we recognize this as an ellipse with semiaxes equal to the arithmetic and geometric means of R and r.

We know how to find the area inside an ellipse from its semiaxes, so to find the volume we just need the altitude. By proportionality,

$$h = \frac{\rho(c)}{R}H = \frac{a^2 - c^2}{aR}H = \frac{r}{a}H,$$

and our volume is

$$V = \frac{1}{3}(\pi a\sqrt{Rr})h = \frac{1}{3}\pi Hr\sqrt{Rr}.$$

A nicer form for this equation is $V = \sqrt{V_0 V_1}$, where V_0 is the volume of the original right circular cone and V_1 is the volume of a similar cone of radius r. That looks like the sort of formula Archimedes might have discovered, but it doesn't seem to be commonly known. Moreover, it just begs for a generalization. Most of the assumptions we made while discovering it are indeed irrelevant; we'll summarize what is really needed in a formal theorem. Its proof is almost immediate.

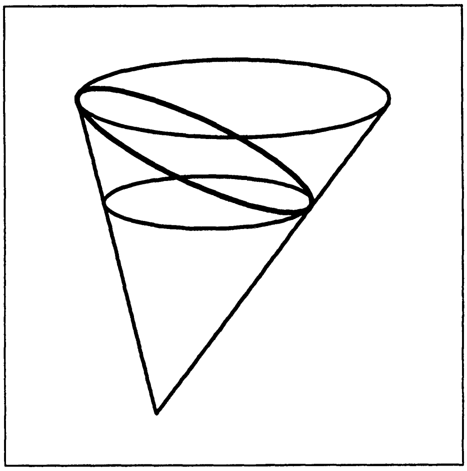


Figure 2. Cones with $V = \sqrt{V_0 V_1}$.

Theorem. *Let C be any truncated cone with an elliptical base, and let C_0 and C_1 be two truncated cones formed from the same conic surface as C by distinct parallel planes tangent to the base of C. Then the volume of C is the geometric mean of the volumes of C_0 and C_1.*

Figure 2 is a sketch illustrating the relationship among the three cones. The conic surface must be the graph of a quadratic equation, so C_0 and C_1 also have elliptical bases.

We can always reduce the theorem to the special case we've already treated. Just choose an affine transformation that maps C_0 and C_1 to right circular cones, and note the equation $V = \sqrt{V_0 V_1}$ is affine invariant.

The assumption that the bases of our cones are elliptical seems to be essential, not just a convenience. For example, the theorem fails if the cones are replaced by pyramids. Perhaps there's a new characterization of ellipses here, just waiting to be discovered.

Department of Mathematical Sciences
New Mexico State University
Las Cruces, New Mexico 88003
rbagby@nmsu.edu

On Regular Bases of Finite Groups

Edward Bertram and Marcel Herzog

Let G denote a finite group, and h a natural number. A subset $B \subseteq G$ is called an *h-basis* of G if

$$B^h = G$$

where $B^h = \{a_1 a_2 \cdots a_h | a_i \in B,\ 1 \le i \le h\}$ is the set of all products of h (not necessarily distinct) elements of G. Clearly, if B is an h-basis of G then $|B|^h \ge |G|$. Every group has the *trivial* h-basis $B = G$.

In 1937, Rohrbach [7, 8] asked whether for *every* $h \ge 2$ there exists a constant $c_h (\ge 1)$ such that for *every* finite group G there exists an h-basis B for G such that $|B| \le c_h |G|^{1/h}$. When $h = 2$, improving the results of Rohrbach [8], and Bertram and Herzog [1] in special cases, Finkelstein et al. [4], and Kozma and Lev [5] independently gave an affirmative answer to Rohrbach's question. Both proofs use the Classification of the Finite Simple Groups.

Extending the definitions in [1] and [5], and following Rohrbach's question, Kozma and Lev [6] call a family $\mathscr{F}$ of finite groups *well-h-based* if there exists a constant c_h such that every $G \in \mathscr{F}$ has an h-basis $B \subseteq G$ satisfying $|B| \le c_h |G|^{1/h}$.

Question 1. Is the family of *all* finite groups well-h-based?

When $h = 2$, an affirmative answer was given in [5] (with $c_2 = 4/\sqrt{3}$), but the answer isn't known in general. Rohrbach showed [8] that the family of all finite cyclic groups is well-h-based for every $h \ge 2$, and Kozma and Lev [6] have recently shown that the family of all finite solvable or alternating groups is well-h-based.

Here, an h-basis U of G will be called *unary* if each element of G is obtained *exactly once* in U^h. This clearly implies that $|U| = |G|^{1/h}$. More generally, an h-basis R of G will be called *regular* if for some natural number r, each element of G is obtained *exactly r times* in R^h. This clearly implies that

$$|R|^h = r|G|.$$

If $G = \{1\}$, the trivial h-basis is a unary basis, and for an arbitrary finite group G the trivial h-basis $R = G$ is regular with $r = |G|^{h-1}$, for all $h \ge 1$.

Question 2. For $h \ge 2$, does there exist a finite group $G, |G| > 1$, with a unary h-basis?

For $h = 2$ the answer has been known for some time, and the authors are aware of two unpublished proofs, due to John Thompson and S. Gagola, both using the representation theory of finite groups, of

Theorem 1. *Only $G = \{1\}$ can have a unary 2-basis.*

Dimovski [3] completely answered Question 2 by proving

Theorem 2. *For each $h \ge 2$, only $G = \{1\}$ can have a unary h-basis.*

On the other hand, Dimovski [3] constructed an infinite group with a unary 2-basis.

Question 3. Does there exist a finite group G with a non-trivial regular h-basis?

Dimovski [3] proved, again using representation theory,

Theorem 3. *If $R \subseteq G$ is a regular 2-basis of G, then $R = G$.*

It was a pleasant surprise when Professor Gary Sherman sent us a preprint [2] by Matthew Cushman, then a senior at Carnegie Mellon, in which Cushman proves Theorem 2 by elementary methods. Applying a variation of Cushman's method, and Fermat's Little Theorem, we can now give a complete answer to Question 3:

Theorem 4. *If $R \subseteq G$ is a regular-h-basis of G, then $R = G$.*

Proof: If $h = 1$ then $R = G$, so we may assume that $h \geq 2$ and $|G| > 1$. Since R is a regular h-basis of G, we must have

$$|R| \leq |G| \quad \text{and} \quad |R|^h = r|G| \tag{1}$$

for some natural number r. It follows that

$$r \leq |R|^{h-1}. \tag{2}$$

Now let p be a prime satisfying

$$p > |G|^{h-1} \tag{3}$$

and note that $p > h$ since we are assuming that $|G| > 1$. We now define

$$X = \left\{(a_1, a_2, \dots, a_p) | a_i \in R \quad a_1 a_2 \cdots a_p = 1\right\}.$$

For each $(p-h)$-tuple $(a_{h+1}, \dots, a_p)$ with $a_i \in R$, there are exactly r h-tuples $(a_1, \dots, a_h)$, $a_i \in R$ such that $(a_1, \dots, a_h, a_{h+1}, \dots, a_p) \in X$, and thus

$$|X| = r|R|^{p-h}. \tag{4}$$

Clearly, if $(a_1, \dots, a_p) \in X$, then $(a_p, a_1, a_2, \dots, a_{p-1}) \in X$. Let the cyclic group $C_p = \langle g \rangle$ of order p act on X, by $(a_1, \dots, a_p)g = (a_p, a_1, a_2, \dots, a_{p-1})$. Then each $\langle g \rangle$-orbit in X is of length 1 or p.

Suppose that all $\langle g \rangle$-orbits in X are of length p. Then $p \mid |X|$ and hence by (4)

$$p \,\Big|\, |rR|^{p-h}. \tag{5}$$

But by (1), (2), and (3) we have $p > |G|^{h-1} \geq |R|^{h-1} \geq r$. Thus $p > |R|$ (since $h \geq 2$) as well as $p > r$, and (5) is impossible.

Thus, there exists a $\langle g \rangle$-orbit in X of length 1, consisting of $(a_1, \dots, a_p)$, say. By the definition of the action of g, we must have $a_1 = a_2 = \cdots = a_p$, so $a_1^p = 1$. But since $p > |G|$ we must have $a_1 = 1$, and hence the $\langle g \rangle$-orbit in X of length 1 is *unique*, consisting of $(1, 1, 1, \dots, 1)$. Thus

$$|X| = r|R|^{p-h} \equiv 1 \pmod{p}. \tag{6}$$

Since $p > |R|, (p, |R|) = 1$, so by Fermat's Little Theorem we have

$$|R|^{p-1} \equiv 1 \pmod{p}. \tag{7}$$

Subtracting (7) from (6), we obtain

$$|R|^{p-h}\left(r - |R|^{h-1}\right) \equiv 0 \pmod{p},$$

and as $p > r, |R|^{h-1}, |R|$, we must have $r = |R|^{h-1}$, and then by (1), $|R| = |G|$ and hence $R = G$ as claimed. ■

As a corollary we have Theorem 2:

Corollary. *If G has a unary h-basis, then either $h = 1$, or $|G| = 1$.*

Proof: By Theorem 4 and its proof, if $h \geq 2$ then $1 = r = |G|^{h-1}$, and thus $|G| = 1$. ■

REFERENCES

1. E. A. Bertram and M. Herzog, On medium-size subgroups and bases of finite groups, *J. Combin. Theory Ser. A* **57** (1991), 1–14.
2. M. Cushman, An elementary proof that finite groups lack unique product structures, *J. Algebra* (to appear).
3. D. Dimovski, Groups with unique product structures, *J. Algebra* **146** (1992), 205–209.
4. L. Finkelstein, D. Kleitman, and T. Leighton, Applying the classification theorem for finite simple groups to minimize pin count in uniform permutation architectures, *Proceedings, Aegean Workshop on Computing*, Lecture Notes in Computer Science, Vol. 319, pp. 247–256, Springer-Verlag, Berlin/New York, 1988.
5. G. Kozma and A. Lev, Bases and decomposition numbers of finite groups, *Arch. Math.* **58** (1992), 417–424.
6. G. Kozma and A. Lev, On h-bases and h-decompositions of the finite solvable and alternating groups, *J. Number Theory* **49** (1994), 385–391.
7. H. Rohrbach, Ein Beitrag zur additiven Zahlentheorie, *Math. Z.* **42** (1937), 1–30.
8. H. Rohrbach, Anwendung eines Satzes der additiven Zahlentheorie auf eine gruppentheoretische Frage, *Math. Z.* **42** (1937), 538–542.

Department of Mathematics
University of Hawaii at Manoa
Honolulu, Hawaii 96822
ed@math.hawaii.edu

School of Mathematical Sciences
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University, Tel-Aviv, Israel
herzog@math.tau.ac.il

A Postscript

P.S. "I want to be a mathematician" because of "the thrills of abstraction". At all times, I have been privileged to know "how to talk mathematics", "how to write mathematics", and "what to publish" in "mathematics as a creative art". Skimming through "the legend of John von Neumann", I caught "a glimpse into Hilbert space". In the "heart of mathematics", there are "some books of auld lang syne" with "problems for mathematicians, young and old". Just to "think it gooder", "I have a photographic memory" of all the "innovation in mathematics" highlighting "American mathematics from 1940 to the day before yesterday".

Composed by Man-Duen Choi for Paul Halmos' eightieth birthday (March 3, 1996). All quoted items are titles of Halmos' books and articles

THE EVOLUTION OF...

Edited by: **Abe Shenitzer**

Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Scientist As Rebel[1]

Freeman Dyson

I have a low opinion of reductionism, which seems to me to be at best irrelevant and at worst misleading as a description of what science is about. Let me begin with pure mathematics. Here the failure of reductionism has been demonstrated by rigorous proof. This will be a familiar story to many of you. The great mathematician David Hilbert, after thirty years of high creative achievement on the frontiers of mathematics, walked into the blind alley of reductionism. In his later years he espoused a programme of formalization, which aimed to reduce the whole of mathematics to a collection of formal statements using a finite alphabet of symbols and a finite set of axioms and rules of inference. This was reductionism in the most literal sense, reducing mathematics to a set of marks written on paper, and deliberately ignoring the context of ideas and applications that give meaning to the marks. Hilbert then proposed to solve the problems of mathematics by finding a general decision-process that could decide, given any formal statement composed of mathematical symbols, whether that statement was true or false. He called the problem of finding this decision-process the *Entscheidungsproblem*. He dreamed of solving the *Entscheidungsproblem* and thereby solving as corollaries all the famous unsolved problems of mathematics. This was to be the crowning achievement of his life, the achievement that would outshine all the achievements of earlier mathematicians who solved problems only one at a time.

The essence of Hilbert's programme was to find a decision process that would operate on symbols in a purely mechanical fashion, without requiring any understanding of their meaning. Since mathematics was reduced to a collection of marks on paper, the decision process should concern itself only with the marks and not with the fallible human intuitions out of which the marks were reduced. In spite of prolonged efforts of Hilbert and his disciples, the *Entscheidungsproblem* was never solved. Success was achieved only in highly restricted domains of mathematics, excluding all the deeper and more interesting concepts. Hilbert never gave up hope, but as the years went by his programme became an exercise in formal logic having little connection with real mathematics. Finally, when Hilbert was seventy

[1]This essay is part (pp. 5–11) of Freeman Dyson's introduction to the book *Nature's Imagination. The Frontiers of Scientific Vision* (ed. John Cornwell), a collection of essays published by Oxford University Press in 1995. This part of the introduction is being reprinted by permission of Oxford University Press. Some readers may be interested in Saunders MacLane's reaction to Dyson's essay, which took the form of a letter to the editors ('A Matter of Temperment') in the October 5, 1995, *New York Review of Books*. MacLane's letter is followed by Dyson's reply; the same issue contains Steven Weinberg's review (pp. 39–42) of the entire book.

years old, Kurt Gödel proved by a brilliant analysis that the *Entscheidungsproblem* as Hilbert formulated it cannot be solved. Gödel proved that in any formalization of mathematics including the rules of ordinary arithmetic a formal process for separating statements into true and false cannot exist. He proved the stronger result which is now known as Gödel's Theorem, that in any formalization of mathematics including the rules of ordinary arithmetic there are meaningful arithmetical statements that cannot be proved true or false. Gödel's Theorem shows conclusively that in pure mathematics reductionism does not work. To decide whether a mathematical statement is true, it is not sufficient to reduce the statement to marks on paper and to study the behaviour of the marks. Except in trivial cases, you can decide the truth of a statement only by studying its meaning and its context in the larger world of mathematical ideas.

It is a curious paradox that several of the greatest and most creative spirits in science, after achieving important discoveries by following their unfettered imaginations, were in their later years obsessed with reductionist philosophy and as a result became sterile. Hilbert was a prime example of this paradox. Einstein was another. Like Hilbert, Einstein did his great work up to the age of forty without any reductionist bias. His crowning achievement, the general relativistic theory of gravitation, grew out of a deep physical understanding of natural processes. Only at the very end of his ten-year struggle to understand gravitation did he reduce the outcome of his understanding to a finite set of field-equations. But like Hilbert, as he grew older he concentrated his attention more and more on the formal properties of his equations, and he lost interest in the wider universe of ideas out of which the equations arose. His last twenty years were spent in a fruitless search for a set of equations that would unify the whole of physics, without paying attention to the rapidly proliferating experimental discoveries that any unified theory would finally have to explain. I do not need to say more about this tragic and well-known story of Einstein's lonely attempt to reduce physics to a finite set of marks on paper. His attempt failed as dismally as Hilbert's attempt to do the same thing with mathematics. I shall instead discuss another aspect of Einstein's later life, an aspect that has received less attention than his quest for the unified field equations: his extraordinary hostility to the idea of black holes.

Black holes were invented by Oppenheimer and Snyder in 1940. Starting from Einstein's theory of general relativity, Oppenheimer and Snyder found solutions of Einstein's equations describing what happens to a massive star when it has exhausted its supplies of nuclear energy. The star collapses gravitationally and disappears from the visible universe, leaving behind only an intense gravitational field to mark its presence. The star remains in a state of permanent free fall, collapsing endlessly inward into the gravitational pit without ever reaching the bottom. This solution of Einstein's equations was profoundly novel. It has had enormous impact on the later development of astrophysics. We now know that black holes ranging in mass from a few suns to a few billion suns actually exist and play a dominant role in the economy of the universe. In my opinion, the black hole is incomparably the most exciting and the most important consequence of general relativity. Black holes are the places in the universe where general relativity is decisive. But Einstein never acknowledged his brainchild. Einstein was not merely sceptical, he was actively hostile to the idea of black holes. He thought that the black-hole solution was a blemish to be removed from his theory by a better mathematical formulation, not a consequence to be tested by observation. He never expressed the slightest enthusiasm for black holes, either as a concept or as a physical possibility. Oddly enough, Oppenheimer too in later life was uninterested

in black holes, although in retrospect we can say that they were his most important contribution to science. The older Einstein and the older Oppenheimer were blind to the mathematical beauty of black holes, and indifferent to the question whether black holes actually exist.

How did this blindness and this indifference come about? I never discussed this question directly with Einstein, but I discussed it several times with Oppenheimer and I believe that Oppenheimer's answer applies equally to Einstein. Oppenheimer in his later years believed that the only problem worthy of the attention of a serious theoretical physicist was the discovery of the fundamental equations of physics. Einstein certainly felt the same way. To discover the right equations was all that mattered. Once you had discovered the right equations, then the study of particular solutions of the equations would be a routine exercise for second-rate physicists or graduate students. In Oppenheimer's view, it would be a waste of his precious time, or of mine, to concern ourselves with the details of particular solutions. This was how the philosophy of reductionism led Oppenheimer and Einstein astray. Since the only purpose of physics was to reduce the world of physical phenomena to a finite set of fundamental equations, the study of particular solutions such as black holes was an undesirable distraction from the central goal. Like Hilbert, they were not content to solve particular problems one at a time. They were entranced by the dream of solving all the basic problems at once. And as a result, they failed in their later years to solve any problems at all.

In the history of science it happens not infrequently that a reductionist approach leads to spectacular success. Frequently the understanding of a complicated system as a whole is impossible without an understanding of its component parts. And sometimes the understanding of a whole field of science is suddenly advanced by the discovery of a single basic equation. Thus it happened that the Schrödinger equation in 1926 and the Dirac equation in 1927 brought a miraculous order into the previously mysterious processes of atomic physics. The equations of Schrödinger and Dirac were triumphs of reductionism. Bewildering complexities of chemistry and physics were reduced to two lines of algebraic symbols. These triumphs were in Oppenheimer's mind when he belittled his own discovery of black holes. Compared with the abstract beauty and simplicity of the Dirac equation, the black-hole solution seemed to him ugly, complicated, and lacking in fundamental significance.

But it happens at least equally often in the history of science that the understanding of the component parts of a composite system is impossible without an understanding of the behaviour of the system as a whole. And it often happens that the understanding of the mathematical nature of an equation is impossible without a detailed understanding of its solutions. The black hole is a case in point. One could say without exaggeration that Einstein's equations of general relativity were understood only at a very superficial level before the discovery of the black hole. During the fifty years since the black hole was invented, a deep mathematical understanding of the geometrical structure of space-time has slowly emerged, with the black-hole solution playing a fundamental role in the structure. The progress of science requires the growth of understanding in both directions, downward from the whole to the parts and upward from the parts to the whole. A reductionist philosophy, arbitrarily proclaiming that the growth of understanding must go only in one direction, makes no scientific sense. Indeed, dogmatic philosophical beliefs of any kind have no place in science.

Science in its everyday practice is much closer to art than to philosophy. When I look at Gödel's proof of his undecidability theorem, I do not see a philosophical

argument. The proof is a soaring piece of architecture, as unique and as lovely as Chartres cathedral. Gödel took Hilbert's formalized axioms of mathematics as his building-blocks and built out of them a lofty structure of ideas into which he could finally insert his undecidable arithmetical statement as the keystone of the arch. The proof is a great work of art. It is a construction, not a reduction. It destroyed Hilbert's dream of reducing all mathematics to a few equations, and replaced it with a greater dream of mathematics as an endlessly growing realm of ideas. Gödel proved that in mathematics the whole is always greater than the sum of the parts. Every formalization of mathematics raises questions that reach beyond the limits of the formalism into unexplored territory.

The black-hole solution of Einstein's equations is also a work of art. The black hole is not as majestic as Gödel's proof, but it has the essential features of a work of art: uniqueness, beauty, and unexpectedness. Oppenheimer and Snyder built out of Einstein's equations a structure that Einstein had never imagined. The idea of matter in permanent free fall was hidden in the equations, but nobody saw it until it was revealed in the Oppenheimer-Snyder solution. On a much more humble level, my own activities as a theoretical physicist have a similar quality. When I am working, I feel myself to be practising a craft rather than following a method. When I did my most important piece of work as a young man, putting together the ideas of Tomonaga, Schwinger, and Feynman to obtain a simplified version of quantum electrodynamics, I had consciously in mind a metaphor to describe what I was doing. The metaphor was bridge-building. Tomonaga and Schwinger had built solid foundations on one side of a river of ignorance, Feynman had built solid foundations on the other side, and my job was to design and build the cantilevers reaching out over the water until they met in the middle. The metaphor was a good one. The bridge that I built is still serviceable and still carrying traffic forty years later. The same metaphor describes well the greater work of unification achieved by Weinberg and Salam when they bridged the gap between electrodynamics and the weak interactions. In each case, after the work of unification is done, the whole stands higher than the parts.

In recent years there has been great dispute among historians of science, some believing that science is driven by social forces, others believing that science transcends social forces and is driven by its own internal logic and by the objective facts of nature. Historians of the first group write social history, those of the second group write intellectual history. Since I believe that scientists should be artists and rebels, obeying their own instincts rather than social demands or philosophical principles, I do not fully agree with either view of history. Nevertheless, scientists should pay attention to the historians. We have much to learn, especially from the social historians.

Many years ago, when I was in Zürich, I went to see the play *The physicists* by the Swiss playwright Dürrenmatt. The characters in the play are grotesque caricatures, wearing the costumes and using the names of Newton, Einstein, and Möbius. The action takes place in a lunatic asylum where the physicists are patients. In the first act they entertain themselves by murdering their nurses, and in the second act they are revealed to be secret agents in the pay of rival intelligence services. I found the play amusing but at the same time irritating. These absurd creatures on the stage had no resemblance at all to any real physicist. I complained about the unreality of the characters to my friend Markus Fierz, a well-known Swiss physicist who came with me to the play. 'But don't you see?', said Fierz, 'The whole point of the play is to show us how we look to the rest of the human race.' Fierz was right. The image of noble and virtuous dedication to truth, the image that scientists have

traditionally presented to the public, is no longer credible. The public, having found out that the traditional image of the scientist as a secular saint is false, has gone to the opposite extreme and imagines us to be irresponsible devils playing with human lives. Dürrenmatt has held up the mirror to us and has shown us the image of ourselves as the public sees us. It is our task now to dispel these fantasies with facts, showing to the public that scientists are neither saints nor devils but human beings sharing the common weaknesses of our species.

Historians who believe in the transcendence of science have portrayed scientists as living in a transcendent world of the intellect, superior to the transient, corruptible, mundane realities of the social world. Any scientist who claims to follow such exalted ideals is easily held up to ridicule as a pious fraud. We all know that scientists, like television evangelists and politicians, are not immune to the corrupting influences of power and money. Much of the history of science, like the history of religion, is a history of struggles driven by power and money. And yet, this is not the whole story. Genuine saints occasionally play an important role, both in religion and in science. Einstein was an important figure in the history of science, and he was a firm believer in transcendence. For Einstein, science as a way of escape from mundane reality was no pretence. For many scientists less divinely gifted than Einstein, the chief reward for being a scientist is not the power and the money but the chance of catching a glimpse of the transcendent beauty of nature.

Both in science and in history there is room for a variety of styles and purposes. There is no necessary contradiction between the transcendence of science and the realities of social history. One may believe that in science nature will ultimately have the last word, and still recognize an enormous role for human vainglory and viciousness in the practice of science before the last word is spoken. One may believe that the historian's job is to expose the hidden influences of power and money, and still recognize that the laws of nature cannot be bent and cannot be corrupted by power and money. To my mind, the history of science is most illuminating when the frailties of human actors are put into juxtaposition with the transcendence of nature's laws.

Francis Crick is one of the great scientists of our century. He has recently published his personal narrative of the microbiological revolution that he helped to bring about, with a title borrowed from Keats, *What mad pursuit*. One of the most illuminating passages in his account compares two discoveries in which he was involved. One was the discovery of the double-helix structure of DNA, the other was the discovery of the triple-helix structure of the collagen molecule. Both molecules are biologically important, DNA being the carrier of genetic information, collagen being the protein that holds human bodies together. The two discoveries involved similar scientific techniques and aroused similar competitive passions in the scientists racing to be the first to find the structure. Crick says that the two discoveries caused him equal excitement and equal pleasure at the time he was working on them. From the point of view of a historian who believes that science is a purely social construction, the two discoveries should have been equally significant. But in history as Crick experienced it, the two helices were not equal. The double helix became the driving-force of a new science, while the triple helix remained a footnote of interest only to specialists. Crick asks the question, How are the different fates of the two helices to be explained? He answers the question by saying that human and social influences cannot explain the difference, that only the transcendent beauty of the double helix structure and its genetic function can explain the difference. Nature herself, and not the scientist, decided

what was important. In the history of the double helix, transcendence was real. Crick gives himself the credit for choosing an important problem to work on, but, he says, only nature herself could tell how transcendentally important it would turn out to be.

My message is that science is a human activity, and the best way to understand it is to understand the individual human beings who practise it. Science is an art form and not a philosophical method. The great advances in science usually result from new tools rather than from new doctrines. If we try to squeeze science into a single philosophical viewpoint such as reductionism, we are like Procrustes chopping off the feet of his guests when they do not fit on to his bed. Science flourishes best when it uses freely all the tools at hand, unconstrained by preconceived notions of what science ought to be. Every time we introduce a new tool, it always leads to new and unexpected discoveries, because Nature's imagination is richer than ours.

Lester R. Ford Awards for 1995

Winners of the Lester R. Ford Awards for expository papers appearing in Volume 102 (1995) of the *Monthly* are:

Martin Aigner, Turán's Graph Theorem, 808–816.
Sheldon Axler, Down with Determinants!, 139–154.
John Oprea, Geometry and the Foucault Pendulum, 515–522.

The Lester R. Ford awards, established in 1964, are made to authors of expository articles published in the *Monthly*. The Awards are named for Lester R. Ford, Sr., a distinguished mathematician, editor of the *Monthly* (1942–1946), and President of the Mathematical Association of America (1947–1948).

THE AUTHORS

HAROLD G. DIAMOND was introduced to the Prime Number Theorem in a class of Prof. W. H. J. Fuchs at Cornell University. Prime numbers have become a lifelong fascination and the focus of his research. He wrote a Ph.D. thesis on the subject at Stanford University under Paul J. Cohen in 1965, and profited from student and post doctoral contact with Paul Turán, K. Chandrasekharan, A. Beurling, and A. Selberg. Since 1967, Diamond has been a faculty member at the University of Illinois.

PAUL T. BATEMAN obtained three degrees from the University of Pennsylvania, writing his Ph.D. thesis in 1946 under Hans Rademacher. From 1950 until his retirement in 1989 he was on the faculty of the University of Illinois at Urbana-Champaign. With Harold Diamond, he was a coeditor of the Problem Section of the MONTHLY from 1986 through 1991. While a graduate student he had the good fortune to sit next to Jacques Hadamard at a 1942 sectional meeting of the American Mathematical Society held at Columbia University.

PETER GRITZMANN was born in 1954 in Dortmund, Germany, and is a Professor of Mathematics at the University of Trier. His diploma degree was from the University of Dortmund in 1980, his Ph.D. and Habilitation from the University of Siegen in 1980 and 1984. In addition to various appointments in Germany, he has on several occasions been a visiting professor at the University of Washington in Seattle, and in 1987 he was a senior research fellow at the Institute for Mathematics and its Applications in Minneapolis. Awards received include a Feodor Lynen Fellowship of the Humboldt Foundation and, jointly with V. Klee, a Max-Planck Research Award. Gritzmann's current main research interest is the field of Computational Convexity, which is the algorithmic study of high-dimensional convex bodies (especially polytopes) with a view to applying the resulting knowledge to bodies that arise in other mathematical disciplines or in the mathematical modeling of practical problems from outside mathematics.

THOMAS BURGER was born in 1964 in Düsseldorf, Germany. He received an M.S. degree from the University of Massachusetts in 1988 and a diploma degree from the University of Freiburg in 1990. After three years of work in the software industry, he found that he missed the fine and elegant ways of mathematics, and decided to return to his studies. He is now at the University of Trier, struggling with his Ph.D. thesis, which will be on Computational Convexity.

VICTOR KLEE was born in 1925 in San Francisco, received a B.A. from Pomona College in 1945 and a Ph.D. from the University of Virginia in 1949. He has been based at the University of Washington since 1953, but over the years has "done time" at several other universities and research institutes. He was President of the MAA for two years in the early 1970's, and has received the MAA's Award for Distinguished Service, its L. R. Ford Award, and its C. B. Allendoerfer Award. Most of his research activities have in some way involved convex geometry, even though the end results have sometimes been classified as functional analysis, topology, optimization, or combinatorics. His primary current interest is Computational Convexity.

ALEKSANDAR JURIŠIĆ (if the surname was written as Yureesheech, this could help to guess its pronunciation) received a B.A. from University of Ljubljana in '87 (working under Jože Vrabec on applications of topology in combinatorics), and M.Sc., Ph.D. from University of Waterloo in '90, '95 (working under Chris Godsil in the field of algebraic combinatorics). He now has an industrial postdoctoral position at Certicom Corp. and University of Waterloo, working in cryptography and algorithmic number theory. His main research interests are discrete mathematics and geometry. He

loves problem solving and trying to make difficult things look easy. In his free time he enjoys playing basketball and teaching recreational mathematics. He wrote a problems book for mathematics competitions in '89.

ARTHUR WHITE received his degrees from Oberlin College in 1961 and Michigan State University in 1969, with actuarial science and military service intervening. He has taught at Western Michigan University since 1969, with one sabbatical leave to Royal Holloway College and two to Oxford University (Wolfson College). A topological graph theorist, he finds that by imbedding the right graph of the right group, with helpful symmetries on the right surface, he is able to compose music for English church bells.

HERBERT I. BROWN received his Ph.D. from Rutgers University and is Associate Professor of Mathematics at The State University of New York at Albany. His research specialization is in the application of functional analysis to the study of sequence spaces and summability theory. He is Director, Computer Assisted Instruction in Mathematics, and was instrumental in the creation of interactive computerized classrooms within the department of mathematics.

JAMES W. BURGMEIER was born in 1943. He spent 22 years in Houston; two years each in Boulder and Albuquerque; and 27 years in Burlington, VT at the University of Vermont. He is Professor of Mathematics at the University of Vermont. His interests include mathematics, mathematics education, software development, and gardening.

DAVID S. DUMMIT was born in 1954 and has since spent time at locations on both U.S. coasts and points in between, and is Associate Professor of Mathematics at the University of Vermont. Undergraduate years were spent at CalTech earning a B.S. and M.S. in mathematics and graduate years were spent at Princeton earning an M.A. and Ph.D. in mathematics. Research interests are centered on algebraic number theory, arithmetic algebraic geometry, and computational number theory; non-mathematical interests center on his family and dogsledding with his team of Siberian huskies on the trails in the beautiful Vermont countryside.

MIKE EGGAR grew up in the land of the kangaroo,
but has taught at Edinburgh University since 1972.
He can scarce believe he has been here so long—
A good sign? Or is that wrong?
At Sydney and Oxford he studied, and a topologist became,
but neither university is for his poetry to blame.

WILLIAM MASSEY was born in central Illinois and grew up there during the great depression. He obtained his bachelor's and master's degree from the University of Chicago just before World War II, and served in the U.S. Navy for four years during the war. After the war he got his Ph.D. at Princeton. After ten years on the faculty at Brown University and thirty one years at Yale, he retired with the title Professor Emeritus in 1991. His four books and most of his research papers have been concerned with algebraic topology. Retirement has permitted him to spend more time on his favorite hobby, bird watching.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Two copies of all items needed to evaluate the problem should be sent. A third copy of the the problem and solution is often useful; please include one if possible.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before April 30, 1997. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10550. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.*

What rational values are possible for $\cos\phi/\cos\theta$ when ϕ and θ are rational multiples of π?

10551. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.*

Suppose that $f(x, y)$ is continuous and nonnegative for $x^2 + y^2 \leq 1$, that $f_x(x, y)$ and $f_y(x, y)$ are continuous for $x^2 + y^2 < 1$, and that $f(0, 0) = f(1, 0) = 0$. Must there be a point (x, y) with $0 < x^2 + y^2 < 1$ where $f_x(x, y) = f_y(x, y) = 0$?

10552. *Proposed by Daniel Goffinet, Saint Étienne, France, and Michel Quercia, Dijon, France.*

Is it possible to find a function $f: \mathbb{R} \to \mathbb{R}$ whose graph has exactly three axes of symmetry?

10553. *Proposed by Bjorn Poonen, Mathematical Sciences Research Institute, Berkeley, CA, Jim Propp, Massachusetts Institute of Technology, Cambridge, MA, and Richard Stong, Rice University, Houston, TX.*

Say that a sequence $\langle q\rangle = q_0, q_1, q_2, \ldots$ of integers has the *divisible differences property* if $(n-m)|(q_n - q_m)$ for all n and m.

(**a**) Show that if $\langle q\rangle$ has the divisible differences property and $\limsup |q_n|^{1/n} < e-1$, then there is a polynomial $Q(x)$ such that $q_n = Q(n)$.

(**b**) Show that there is a sequence $\langle q\rangle$ that has the divisible differences property and satisfies $\limsup |q_n|^{1/n} \leq e$, for which q_n is not given by a polynomial in n.

(**c**)* Is it true that $\limsup |q_n|^{1/n} \geq e$ for all non-polynomial $\langle q\rangle$ with the divisible differences property?

10554. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.*

Suppose that the functions $f(r)$ and $g(r)$ are positive and continuous for $0 \leq r < 1$. Using polar coordinates (r, θ), define a metric on the open unit disk by the formula

$$ds^2 = f(r)^2 dr^2 + g(r)^2 r^2 d\theta^2.$$

For what choices of $f(r)$ and $g(r)$ does the open unit disk with this metric furnish a model of the hyperbolic plane?

10555. *Proposed by Jeffrey C. Lagarias, AT&T Research, Murray Hill, NJ, and Eric M. Rains, Harvard University, Cambridge, MA.*

(**a**) Suppose that $f(t) \in \mathbb{R}[t]$ is a polynomial that maps rationals to rationals and irrationals to irrationals. Show that $f(t) = at + b$ with a and b rational.

(**b**) Does the same conclusion hold under the weaker assumption that $f: \mathbb{R} \to \mathbb{R}$ is an algebraic function (i.e., if there is a polynomial $P(x, y) \in \mathbb{R}[x, y]$ such that $P\big(t, f(t)\big)$ is identically zero)?

10556. *Proposed by Joseph Lewittes, Herbert H. Lehman College, Bronx, NY.*

A real quadratic irrational α is the root of a quadratic equation $ax^2 + bx + c = 0$ with integer coefficients and discriminant $d = b^2 - 4ac > 0$. We also require that $a > 0$ and $\gcd(a, b, c) = 1$ to obtain well-defined coefficients and a discriminant that is a function of α. The continued fraction of α, $[a_0, a_1, \ldots]$, is obtained by setting $\alpha_0 = \alpha$ and recursively defining $a_i = \lfloor \alpha_i \rfloor$ and $\alpha_{i+1} = 1/(\alpha_i - a_i)$. As part of the proof that the continued fraction of quadratic irrationals are eventually periodic, Lagrange showed that all α_i have the same discriminant. However, numbers having the same discriminant may have different periods.

Let $e(\alpha)$ be the number of a_i in the period of the continued fraction of α that are even. If α and α' have the same discriminant, show that $e(\alpha) \equiv e(\alpha') \pmod 2$.

NOTES

(**10550**) These pages have already expressed our sadness at the death of Raphael M. Robinson (see [1995, 469]). The present editors remember him as an active and enthusiastic supporter of this Problem Section. In honor of the eighty-fifth anniversary of his birth on November 2, we have collected his contributions that had been chosen by the usual process, together with solutions to problems that he had proposed. Appropriate new problems by other authors were chosen to add to this collection. (**10555**) A more precise statement of the hypothesis in (**b**)

is that $f: \mathbb{R} \to \mathbb{R}$ forms a single real-analytic branch of an algebraic function (**10556**) With $d = 204$, the continued fraction of $\alpha = 7 + \sqrt{51}$ is purely periodic with period $\langle 14, 7 \rangle$ (and $e = 1$), while the continued fraction of $\alpha' = (\sqrt{51} + 6)/3$ is purely periodic with period $\langle 4, 2, 1, 1, 1, 2 \rangle$ (and $e = 3$); with $d = 321$, the continued fraction of $\beta = (\sqrt{321} + 17)/2$ is purely periodic with period $\langle 17, 2, 5, 2 \rangle$ (and $e = 2$), while the continued fraction of $\beta' = (\sqrt{321} + 9)/24$ is purely periodic with period $\langle 1, 8, 4, 2, 1, 2 \rangle$ (and $e = 4$).

SOLUTIONS

Enumerating admissible graphs on a 3-torus

10278 [1993, 76]. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.*

A three-dimensional torus is formed from Euclidean 3-space by reducing each coordinate modulo 3. An *admissible* graph on this torus is one whose vertices are the 27 lattice points and whose edges are unit segments chosen so that exactly two meet at each vertex and form a right angle there. Find the total number of admissible graphs, taking account of position as well as of form.

Solution by A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. The total number of such graphs is 10752.

We introduce convenient terminology for special vertex subsets.

A *line* of the torus is a set of three collinear points whose coordinates are equal in two positions. In an admissible graph G, the subgraph induced by any line has at most one edge. Since there are 27 lines and the graph contains exactly 27 edges, each line induces exactly one edge of G.

A *layer* is a set of nine points with fixed third coordinate. The torus has three layers, each layer containing six lines. Since a layer L contains six lines, it must contain exactly 6 edges in G. These six edges have 12 endpoints. Each vertex in L is the endpoint of at least one and at most two edges in L. Hence three vertices of L have two neighbors in L, and six vertices of L have one neighbor in L and one in another layer.

Vertices with two neighbors in their own layer will be called *generating points*. Each vertical line must contain an edge of G and hence contains at most one generating point. Two generating points on the same line in the same layer must be adjacent in G, so a line perpendicular to the vertical cannot contain three generating points.

We count the total number of configurations of nine generating points meeting these conditions.

The first layer has either (1) a generating point neighboring two other generating points, (2) two generating points which are neighbors and one which is isolated, or (3) three isolated generating points, as illustrated below:

$$\begin{array}{ccc} . & . & . \\ \times & . & . \\ \times & \times & . \end{array} \qquad \begin{array}{ccc} . & \times & . \\ \times & . & . \\ \times & . & . \end{array} \qquad \begin{array}{ccc} . & . & \times \\ . & \times & . \\ \times & . & . \end{array}$$

There are 36 ways to choose a layer of type 1, 36 others of type 2, and 6 of type 3. If the first layer is given, the second layer must have three generators chosen from the remaining six positions, and the third layer will have generators that must fall in the final three positions.

Since the three generators on any level may not fall in a line, this gives us 16 possibilities for the second layer if the first layer is of type 1, 18 if it is of type 2, and 20 if it is of type 3. This gives $36 \cdot 16 + 36 \cdot 18 + 6 \cdot 20 = 1344$ possible configurations of generating points.

These configurations of generating points determine the nine vertical edges, but they do not completely determine the edges lying in the layers. Using the fact that each non-generating point has one neighbor in G within its layer, we can check for each of the three layer types that there are exactly two ways in which a layer can be filled in with edges to make a layer with that configuration of generating points. This gives $2^3 \cdot 1344 = 10752$ candidates for admissible graphs.

Furthermore, every graph constructed as above is admissible. Consider an arbitrary vertex p in one of the graphs we have constructed. If p is a generating point, exactly two edges meet at p and form a right angle there. (The two edges will be in the same layer as p.) If p is any other vertex, it is the endpoint of one edge in the same layer as p and is also adjacent to the other vertex in its vertical line that is not a generating point. In either case, p has the required property.

Editorial comment. The proposer also noted that there are 12 types of admissible graphs under the isometries of the torus. Of these six contain a single cycle of length 27, four contain cycles of lengths 13, 7, 7, and two contain cycles of length 9, 9, 9.

Solved also by S. M. Gagola Jr. and the proposer. One incorrect solution was received.

Partitions of $\{1, \ldots, n\}$ with Equal Sums

10294 [1993, 291]. *Proposed by Derek A. Holton, University of Otago, Dunedin, New Zealand.*

Given a positive integer n, let $\mathbf{N} = \{1, \ldots, n\}$. For a positive integer k, say that n is k-good if $\mathbf{N}$ can be partitioned into k sets each with the same sum. Show that n is k-good if k divides $\binom{n}{2}$ and n is sufficiently large.

Solution by Raphael M. Robinson, University of California, Berkeley, CA. The divisibility condition is misstated in the problem. We prove that n is k-good if and only if k divides $n(n+1)/2$ and $n \geq 2k - 1$. The necessity of these conditions is easy to establish. If each block in the partition has the sum s, then $ks = \sum_{j=1}^{n} j = n(n+1)/2$, which gives the correct divisibility condition. The inequality $n \geq 2k - 1$ holds because there can be at most one block with a single element.

It remains to show the sufficiency of these conditions. We call a set of parameters $\{n, k, s\}$ admissible if $ks = n(n+1)/2$ and $n \geq 2k - 1$. For $n = 2k - 1$, we have $s = 2k - 1$ and must use the partition

$$\{2k-1\}, \{2k-2, 1\}, \{2k-3, 2\}, \ldots, \{k, k-1\}.$$

For $n = 2k$, we have $s = 2k + 1$ and must use the partition

$$\{2k, 1\}, \{2k-1, 2\}, \ldots, \{k+1, k\}.$$

The proof for $n > 2k$ will be by induction on n. For each admissible set of parameters $\{n, k, s\}$, we shall construct a new partition by using an old partition corresponding to some admissible set of parameters $\{n', k', s'\}$ with $n' < n$. The proof will be divided into cases. In each case, the equation $k's' = n'(n'+1)/2$ will be clear from the construction, but we must check that $n' \geq 2k' - 1$.

Case 1: $n \geq 4k - 1$.

Let $n' = n - 2k$, $k' = k$, and $s' = s - (2n - 2k + 1)$. The new blocks will be obtained from the old blocks by adjoining the k pairs

$$\{n, n-2k+1\}, \{n-1, n-2k+2\}, \ldots, \{n-k+1, n-k\}$$

in any order. Here $n' \geq 2k - 1 = 2k' - 1$.

Case 2: $2k < n < 4k - 1$.

In this case, $n/k > 2$ and $(n+1)/k < 4$. Since $s = n(n+1)/(2k)$, we see that $n+1 < s < 2n$. If we set $n' = s - n - 1$, then $0 < n' < n - 1$. We consider two subcases.

Case 2A: s odd.

Introduce $r = (s-1)/2$. Let $k' = k - n + r$ and $s' = s$. As new blocks, we use the old blocks and the $n - r$ pairs

$$\{n, s-n\}, \{n-1, s-n+1\}, \ldots, \{r+1, r\}.$$

Here $n' - 2k' = n - 2k > 0$, so $n' > 2k'$.

Case 2B: s even.

Let $k' = 2k - 2n + s - 1$ and $s' = s/2$. The old blocks and the set $\{s'\}$ are combined in pairs to form $k - n + s'$ new blocks. The other new blocks are the $n - s'$ pairs

$$\{n, s-n\}, \{n-1, s-n+1\}, \ldots, \{s'+1, s'-1\}.$$

A straightforward calculation shows that $2k(n' - 2k') = (n-2k)(4k-1-n) > 0$. Thus we again have $n' > 2k'$.

Solved also by V. Božin (student, Yugoslavia), R. Holzsager, L. E. Mattics, A. D. Melas (Greece), F. Schmidt, R. Stong, A. A. Tarabay (Lebanon), H. V. Vu (student, Hungary), A. N. 't Woord (The Netherlands), Con Amore Problem Group (Denmark). A solution by R. Guy & J. Selfridge accompanied the original proposal. Three incorrect or incomplete solutions were received.

A Symmetric Factor of a Symmetric Polynomial

10306 [1993, 498]. *Proposed by Seung-Jin Bang, Seoul, Korea.*

Find all positive integers n such that the polynomial

$$a^n(b-c) + b^n(c-a) + c^n(a-b)$$

has $a^2 + b^2 + c^2 + ab + bc + ca$ as a factor.

Solution by Raphael M. Robinson, University of California, Berkeley, CA. This happens only for $n = 1$ and $n = 4$. The quotients in these two cases are 0 and $(a-b)(a-c)(b-c)$. To prove the negative results, we specialize by putting $b = 2$ and $c = 1$. We then ask whether the polynomial $P(a) = a^n - (2^n - 1)a + (2^n - 2)$ can have $Q(a) = a^2 + 3a + 7$ as a factor. The roots of $Q(a)$ are $\frac{-3}{2} \pm \frac{\sqrt{19}i}{2}$, which have magnitude $\sqrt{7}$. If $n \geq 5$ and α is one of the roots of $Q(a)$, we have

$$|P(\alpha)| = \left|\alpha^n - (2^n - 1)\alpha + 2^n - 2\right| \geq \left|\alpha^n\right| - 2^n\big(|\alpha| + 1\big) - 5 > 7^{n/2} - 4 \cdot 2^n > 0.$$

Hence, $P(a)$ cannot have $Q(a)$ as a factor if $n \geq 5$. If $n = 2$ or $n = 3$, then $P(\alpha)$ is also not equal to zero, so the same conclusion is obvious.

Editorial comment. The key step in most solutions was to specialize to zeros of the proposed divisor and to show, in some fashion, that these were not zeros of the dividend.

Solved also by D. Alvis, G. Bhatnagar (student), V. Božin (student, Yugoslavia), A. E. Caicedo Núñez (student, Colombia), R. J. Chapman (U. K.), F. J. Flanigan, J. S. Frame, H. S. Gunaratne (Brunei), R. Holzsager, I. Kastanas, H. K. Krishnapriyan, K.-W. Lau (Hong Kong), W.-Z. Li (China), O. P. Lossers (The Netherlands), T. L. McCoy, A. D. Melas (Greece), I. Nemes (Austria), V. S. Ryko (Russia), F. Schmidt, A. A. Tarabay (Lebanon), M. Vowe (Switzerland), GCHQ Problem Solving Group (U. K.), the MMRS group of Oklahoma State University, National Security Agency Problems Group, PCC Math Problem Solvers Group, and the proposer. Four incorrect or incomplete solutions were received.

Disjoint Curves

10328 [1993, 689]. *Proposed by A. Keith Austin, The University of Sheffield, Sheffield, England.*

Let A and B be sets such that $A \cap B = \emptyset$ and $A \cup B$ is the unit square $[0,1] \times [0,1]$. Prove or disprove the following:

(**a**)* Either there is a function $f: [0,1] \to A$ with $f(0) = (0, y_0)$ for some y_0 and $f(1) = (1, y_1)$ for some y_1, or there is a function $g: [0,1] \to B$ with $g(0) = (x_0, 0)$ for some x_0 and $g(1) = (x_1, 1)$ for some x_1.

(**b**) f and g as in part a cannot both exist.

Solution I of (a) by Fred Galvin, University of Kansas, Lawrence, KS. An easy counterexample may be obtained by taking

$$A = \Big\{ \big(x, 1/2 + 1/2 \sin(1/x)\big) : 0 < x \le 1 \Big\} \cup \Big\{ \big(0, y\big) : 0 \le y \le 1 \Big\}$$
$$B = \big([0,1] \times [0,1]\big) \setminus A.$$

Solution II of (a) by Raphael M. Robinson, University of California, Berkeley, CA. The statement is false. As a counterexample, let A contain boundary points of the unit square with rational coordinates and interior points for which $\max\big(|x - 1/2|, |y - 1/2|\big)$ is rational, and let B contain all other points of the unit square. Then any curve joining two boundary points of the unit square will pass through infinitely many points of A and infinitely many points of B. This is clear if the curve follows the boundary, and if it enters the interior then $\max\big(|x - 1/2|, |y - 1/2|\big)$ will assume both rational and irrational values infinitely often on the curve.

Solution I of (b) by Mark D. Meyerson, U. S. Naval Academy, Annapolis, MD. True. Suppose there were two such functions. Then their graphs are disjoint compact sets (*paths*) and so at a positive distance from each other. Hence, each can be replaced by a polygonal approximation that still satisfies (a). If a polygonal path with distinct endpoints has self-intersections, it can be replaced by a subset with the same endpoints and no self-intersections (an *arc*). Now polygonally connect the endpoints of this graph of the new f outside the square to to make a loop, say by going under the square. This gives a contradiction to the polygonal Jordan curve theorem (which is much more elementary than the full Jordan curve theorem), since the graph of the new g crosses from inside the loop to outside without meeting the loop.

Solution II of (b) by Richard Holzsager, The American University, Washington, DC. Suppose that f and g did exist. Together they would produce a map from the unit square to the plane minus the origin by $H(s,t) = f(s) - g(t)$. Consider what H does on the boundary square. The bottom edge maps into the top half-plane, the right edge maps to the right half-plane, the top edge maps into the bottom half-plane, and the left edge maps into the left half-plane. Such a map has winding number -1. However, H maps the entire square into the complement of the origin, so the winding number must be zero.

Editorial comment. Solution I of (a) uses an example of a set that is connected but not path-connected. Solution II of (a) shows even more than the problem asks, since neither A nor B contains a connected subset meeting the boundary in more than one point. A stronger result of this nature, with a construction using the Axiom of Choice, shows the existence of A and B such that every uncountable closed subset of the unit square contains uncountably many points of both A and B. On the other hand, William Gilbert noted that it is possible for

disjoint connected sets (not *paths*) to join opposite sides of the square. Such constructions can be found in R. J. MacG. Dawson, "Paradoxical connections", this MONTHLY 96 (1989), 31–33 and B. Gelbaum & J. Olmsted, *Counterexamples in Analysis*, Holden-Day, 1964, p. 132. Simeon Stefanov used results found in K. Kuratowski, *Topology, II*, Academic Press, 1968 to study several variations on (a).

Solved also by K. F. Andersen (Canada), J. Boyle, M. Brahm, M. S. Branicky (student), J. L. Bryant, R. J. Chapman (U. K.), W. Gilbert (Canada), J. W. Grossman, K. P. Hart (The Netherlands), G. L. Isaacs, U. Klein (student, Germany), B. Krötz (Germany), M. Lloyd ~& F. Worth, O. P. Lossers (The Netherlands), R. Mabry, J.-H. Mai & F.-P. Zeng (China), R. Martin (student), M. McKinzie, L. F. Meyers, A. Nijenhuis, B. N. Parsons, A. Riese, J. M. Shaw, S. T. Stefanov (Bulgaria), W. Stöcher & T. Wilson (Austria), H. Tamvakis (student), and M. Zerner (Germany). Part (a) only was solved by D. Burke, S. Northshield, A. N. 't Woord (The Netherlands), Aardvark Problem Solving Group, and the Anchorage Math Solutions Group. Part (b) only was solved by the proposer.

Totally Irrational Cantor Translates

10366 [1994, 176]. *Proposed by Raphael M. Robinson, University of California, Berkeley, CA.*

Let $\mathbf{C}$ denote the Cantor set. If s is a real number, by $\mathbf{C} + s$ is meant the set of all sums $c + s$ where $c \in \mathbf{C}$. Find a value of s such that $\mathbf{C} + s$ contains no rational number. This value of s must be given explicitly; an existence proof does not suffice.

Composite solution by Eugene A. Herman, Grinnell College, Grinnell, IA, and O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Whenever digits of a number are mentioned, the base is supposed to be three. By the "n-th digit" we mean the n-th digit after the point, that is the coefficient of 3^{-n}. It is well known that $\mathbf{C}$ consists exactly of the numbers that can be written with just the digits 0 and 2. Now take

$$s = -\sum_{n=1}^{\infty} 3^{-n^2}$$

and suppose that r is rational such that $r - s \in \mathbf{C}$.

If r has a terminating ternary expansion, then for large enough n, the n^2-th digit of $r - s$ is 1, and hence $r - s \notin \mathbf{C}$.

Let the ternary expansion be periodic with period p after the N-th place and choose $n = \max\left(p, \left\lceil \sqrt{N} \right\rceil\right)$. Suppose that, for some k with $n^2 < k < (n+1)^2$, the k-th digit of r is 1 for $n^2 < k < (n+1)^2$. Then, since the k-th digit of $-s$ is 0 and the k-th digit of $r - s$ is not 1, the k-th digit of $r - s$ is 2 and there must be a carry at this place. However, there can then be no further carries after the n^2-th space, so this can only happen once in this block of values of k. However, this block contains at least two full periods of r, so the period of r has no digit equal to 1.

If the $(n+1)^2$-th digit of r is 2, then the $(n+1)^2$-th digit of $r - s$ can only be 0 and a carry must be generated at this place. Thus, for all k with $n^2 < k < (n+1)^2$, the k-th digit of r is 2, the k-th digit of $r - s$ is 0 and a carry is generated. Hence every digit of the period of r is 2, but then r would be equal to a number with a terminating expansion, which has been shown to be impossible.

If the $(n+1)^2$-th digit of r is 0, then the $(n+1)^2$-th digit of $r - s$ can only be 2 and a carry from the next place is required. Thus, for all k with $(n+1)^2 < k < (n+2)^2$, the k-th digit of r is 2, the k-th digit of $r - s$ is 0 and a carry from the next place is required. Again, every digit of the period of r is 2, which is impossible.

Solution by Robert B. Israel, University of British Columbia, Vancouver, B. C., Canada. A suitable s is the number expressed in base 3 as

$$0.(0)(1)(2)(10)(11)(12)(20)(21)(22)(100)\ldots$$

(with the ternary digit of each natural number written one after another; the parentheses are used for the sake of clarity).

The numbers $3^n s \bmod 1$ (i.e., $t_n \in [0, 1)$ such that $3^n s - t_n$ is an integer) are dense in $[0, 1)$, since all possible m-tuples of ternary digits occur. On the other hand, if r is rational, $3^n r \bmod 1$ takes only a finite number of values, while if $c \in \mathbf{C}$, $3^n c \bmod 1 \in \mathbf{C}$. Thus

$$\{ 3^n(r - c) \bmod 1 : n \in \mathbb{N} \}$$

is contained in a finite number of translates of the Cantor set, and is not dense in $[0, 1)$. Therefore s cannot be written as $r - c$, i.e., $\mathbf{C} + s$ contains no rationals.

Solved also by R. Barbara (Lebanon), D. Callan, N. Felsinger, J. W. Grossman, R. Holzsager, F.-A. Izadi (Iran), C. C. Leary, J. H. Lindsey II, R. Mabry, M. D. Meyerson, B. Mixon, F. Richman, F. Schmidt, R. B. Tucker, J. Vinson (student), T. White, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, The Citadel Problem Solving Group, Western Maryland College Problems group, and the proposer.

Bolyai Squares

10374 [1994, 274]. *Proposed by David L. Book, University of Maryland, College Park, MD.*

Given an integer N, characterize the smallest square in the plane containing N lattice points.

Solution by Raphael M. Robinson, University of California, Berkeley, CA. For any given value of N, the desired square can be found by a finite computation described below. Let the *length* of a square be the length of its sides. Let s be the minimal length of a closed square containing N lattice points. If t is an integer, then a square of length t with corners at $(0, 0)$ and (t, t) contains $(t + 1)^2$ lattice points. If we choose t such that $t^2 < N \le (t + 1)^2$, then $s \le t < \sqrt{N}$.

A minimal square containing N lattice points is called *full* if it contains as many lattice points as possible. For example, consider $N = 6$, which yields $s = 2$. Some squares of length 2 contain six lattice points and some contain nine; only the latter type are full minimal squares for $N = 6$.

Every square has the form $\{(x, y) : c_0 \le ax + by \le c_1,\ d_0 \le bx - ay \le d_1\}$. If the square has length s, we write $h = s\sqrt{a^2 + b^2}$, and then $c_1 - c_0 = d_1 - d_0 = h$. We will show that for a full minimal square we may assume that a, b, c_0, c_1, d_0, d_1 are integers with $a^2 + b^2 \le 4s^2$. This makes the computation of s a finite search for each N. With a little more work, one can assure that a, b, c_0, c_1, d_0, d_1 are divisible by $\gcd(a, b)$, so that we may restrict to the case in which a and b are relatively prime.

Two minimal squares for a given N will be called equivalent if one can be obtained from the other by a rigid motion, or a reflection, taking lattice points to lattice points. Every full minimal square can be taken by rotation or reflection to one where the sides $ax + by = c_i$ have slope between 0 and -1, yielding $0 \le a \le b$. This square is equivalent by translation to one whose defining inequalities are $c_0 \le ax + by \le c_0 + h$ and $0 \le bx - ay \le h$, with $0 \le c_0 < h$.

By using a computer program to count the lattice points in all such squares with length $s \le 9$, we found all full minimal squares for $N \le 100$. For each N, the values of a and b and the area s^2 of the minimal squares appear in Table 10374; all areas are exact.

For each $N \le 100$, there is a full minimal square with $c_0 = 0$. The list of inequivalent full minimal squares with $N \le 100$ is completed by adjoining the cases $N = 26$, $c_0 = 3$ and $N = 53$, $c_0 \in \{3, 6\}$. The square for $N = 53$ and $c_0 = 6$ appears in Figure 10374. It is the only full minimal square for $N \le 100$ whose center is a lattice point but has no corner at a lattice point. A 90° rotation about the center preserves both the square and the

N	a	b	s^2	N	a	b	s^2	N	a	b	s^2
1	0	1	0	19-25	0	1	16	53	1	3	48.4
2	1	1	0.5	26	1	3	22.5	54-64	0	1	49
3-4	0	1	1	27-29	1	2	24.2	65	1	2	57.8
5	1	1	2	30-32	1	1	24.5	66-72	1	1	60.5
6-9	0	1	4	33-36	0	1	25	73-81	0	1	64
10	1	2	7.2	37-41	1	1	32	82-85	1	1	72
11-13	1	1	8	42-49	0	1	36	86-89	1	2	80
14-16	0	1	9	50	1	1	40.5	90-100	0	1	81
17-18	1	1	12.5	51-52	1	2	45				

Table 10374

integer lattice. The lattice points on the lower side are $(3, 1)$ and $(6, 0)$, which lie on the line $x + 3y = 6$.

We now prove the previously stated result that a, b, c_0, c_1, d_0, d_1 can be taken to be integers with $a^2 + b^2 \leq 4s^2$. We begin by studying the lattice points on the boundary of a minimal square. Lattice points at vertices will be counted on both incident sides. Some pair of opposite sides must both have lattice points, else we can shrink the square without losing lattice points.

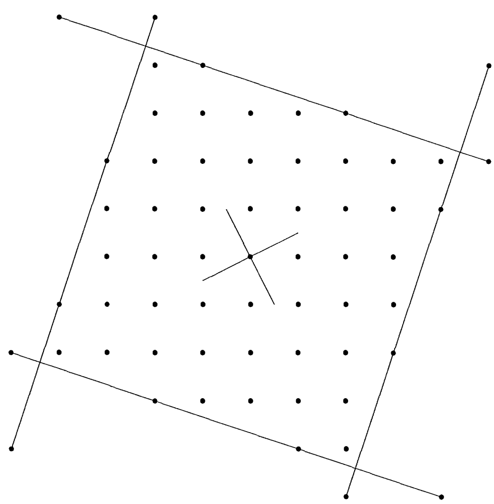


Figure 10374

We claim that some side has at least two lattice points or each side has one. Otherwise, translating the square away from a side with no lattice points yields a square in which one pair of sides have no lattice points and the other pair has one each. Rotating the square around the midpoint of the segment between these two points brings them inside, and then the square may be shrunk.

This also shows that a full minimal square must have lattice points on at least three sides. Otherwise, a translation could increase the number of lattice points in the square.

Now suppose that some side has two lattice points (x_1, y_1) and (x_2, y_2); we may assume

that this side is $ax+by=c_0$. Since $a(x_2-x_1)+b(y_2-y_1)=0$, we may take $a=y_2-y_1$ and $b=x_1-x_2$; these are integers that are not both zero. Since three sides have lattice points, at least three of the numbers c_0, c_1, d_0, d_1 are integers; since $c_1-c_0=d_1-d_0$, the fourth is also an integer. Since the two original points lie on the same side, we have $a^2+b^2 \le s^2$ in this case.

We are left with the case where each side has exactly one lattice point. If the segments joining the lattice points on opposite sides are perpendicular, then a suitable rotation around the intersection point takes the lattice points inside and enables the square to be shrunk. For $i \in \{1,2\}$, let (x_i, y_i) be the lattice point on $ax+by=c_i$, and let (u_i, v_i) be the lattice point on $bx-ay=d_i$. For each variable w, let $\Delta w = w_2 - w_1$. Subtracting the equations for lattice points on opposite sides yields $a\Delta x + b\Delta y = h = b\Delta u - a\Delta v$.

Thus $a(\Delta x+\Delta v)+b(\Delta y-\Delta u)=0$, and we may take $a=\Delta u-\Delta y$ and $b=\Delta x+\Delta v$. These integers cannot both equal 0, since this would make $(\Delta u, \Delta v)$ perpendicular to $(\Delta x, \Delta y)$.

To bound a^2+b^2, we consider the vector $(b,-a)=(\Delta x, \Delta y)-(-\Delta v, \Delta u)$. Since $(\Delta u, \Delta v)$ reaches from the side $bx-ay=d_0$ to the side $bx-ay=d_1$, the rotation $(-\Delta v, \Delta u)$ reaches from the side $ax+by=c_0$ to the side $ax+by=c_1$, as does $(\Delta x, \Delta y)$. Hence the difference $(b,-a)$ is parallel to $ax+by=c_0$ and has length at most $2s$. This yields $a^2+b^2 \le 4s^2$. Furthermore, the numbers c_0, c_1, d_0, d_1 must be integers.

Editorial comment. Prof. Robinson mentioned MONTHLY problem E 1954 [1967, 77; 1968, 545] in which it is shown for all real s that a square of length s contains at most $(s+1)^2$ lattice points, with equality only when s is an integer and the sides are vertical and horizontal. Also, Raphael M. Robinson, "Numbers having m small mth roots mod p,", *Math. Comp.* 61 (1993), 393–413 deals with a similar problem restricted to squares with center at the origin. The number of lattice points then always has the form $N=4i+1$. He remarks that the minimal area is then 12.8 when $N=17$, but Table 10374 indicates that in the unrestricted problem the area goes down to 12.5 and the square contains an extra lattice point.

The proposer did not have a correct solution. In particular, his solution did not allow the square shown in Figure 10374. However, he did give exact formulas for the number of lattice points in a square for which at least two vertices are lattice points. These formulas were sufficient to indicate that a procedure of the type described above would be required to solve the problem for an arbitrary value of N.

One additional incorrect solution was received.

Extreme Values of a Function

10392 [1994, 575]. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Determine the extreme values of

$$\frac{1}{1+x+u}+\frac{1}{1+y+v}+\frac{1}{1+z+w}$$

where $xyz=a^3$, $uvw=b^3$, and $x, y, z, u, v, w > 0$.

Solution by Raphael M. Robinson, University of California, Berkeley, CA. We must assume that a and b are positive. Let S denote the sum of the three fractions. We show that: (1) the greatest lower bound for S is 0; (2) if $a+b \le 1/2$, then S attains a maximum value of $3/(1+a+b)$; and (3) if $a+b > 1/2$, then the least upper bound for S is 2 and this value is not attained.

In our analysis, we shall allow boundary values of 0 and ∞ for the variables if they can be approached by finite positive values satisfying the hypotheses. If (x, y, z) is a boundary triple, then one of the variables must be 0 and one must be ∞, the third being arbitrary. The minimum S on the boundary is 0, attained for $x = y = w = \infty$, $z = u = 0$, and this is the overall minimum. The maximum S on the boundary is 2, attained for $x = y = u = v = 0$, $z = w = \infty$. It remains to be seen whether larger values of S can occur in the interior.

Suppose that the maximum S occurs at an interior point (x, y, z, u, v, w). Hold z, u, v, w fixed and vary x and y. Since xy is then constant, we see that $dy/dx = -y/x$. Hence

$$\frac{dS}{dx} = -\frac{1}{(1+x+u)^2} + \frac{1}{(1+y+v)^2} \cdot \frac{y}{x}.$$

This must vanish, so that

$$x(1+y+v)^2 = y(1+x+u)^2. \tag{$*$}$$

A similar argument shows that

$$u(1+y+v)^2 = v(1+x+u)^2.$$

Thus $u/x = v/y$, and these must also equal w/z. So we may put

$$u = hx, \quad v = hy, \quad w = hz,$$

where $h = b/a$. Condition $(*)$ becomes

$$x\big(1 + (1+h)y\big)^2 = y\big(1 + (1+h)x\big)^2,$$

which reduces to $x - y = (1+h)^2 xy(x-y)$. Hence

$$x = y \quad \text{or} \quad xy = (1+h)^{-2}.$$

Similarly,

$$x = z \quad \text{or} \quad xz = (1+h)^{-2}$$

and

$$y = z \quad \text{or} \quad yz = (1+h)^{-2}.$$

If we do not have $x = y = z$, then, by permuting the variables if necessary, we may assume that $x = y$, $xz = yz = (1+h)^{-2}$, hence

$$x = y = (1+h)^2 a^3, \quad z = (1+h)^{-4} a^{-3},$$

so that

$$x + u = y + v = (a+b)^3, \quad z + w = (a+b)^{-3}.$$

These yield

$$S = \frac{2 + (a+b)^3}{1 + (a+b)^3}.$$

Since this value is less than 2, it cannot be the maximum for S. The only remaining possibility is that $x = y = z = a$, $u = v = w = b$, hence $S = 3/(1+a+b)$. This furnishes the maximum for S if it is at least 2 (i.e., if $a + b \leq 1/2$).

Solved also by R. Holzsager, J. H. Lindsey II, O. P. Lossers (The Netherlands), J. Merickel, the New Mexico Tech Problem Solving Group, National Security Agency Problems Group, and the proposer. Three incomplete and three incorrect solutions were received.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington IN 47405

Algebraic Topology: A First Course. By William Fulton, Springer-Verlag, 1995, (Graduate Texts in Mathematics no. 153), xvii + 430, $39.50

Reviewed by **William S. Massey**

This textbook grew out of undergraduate courses the author taught at Brown University and the University of Chicago. The author states that the book is designed for students of mathematics or science who are not aiming to become practicing algebraic topologists without however discouraging those who are budding topologists.

There are already many textbooks for a first course in algebraic topology. It may be a slight exaggeration to say that there is a consensus about the topics that should be taken up in such a course. However, the majority of the current texts seem to be built around the following topics: the fundamental group, covering spaces, the classification of compact 2-manifolds, singular homology and cohomology theory, the use of simplicial or CW-complexes to compute homology groups, and various applications of the techniques developed to actual problems. Of course many of these texts also treat a few other topics, but usually those listed form the core of the book. Given this quasi-consensus, it is natural to fear that any new textbooks on the subject will be only a slight variation on the existing texts. Fortunately, that is not the case for the book under review: Fulton has written a book that *is* different. We will try to make clear how this book is different as we discuss the topics taken up.

Following the standard practice in algebraic topology textbooks on this level, the main emphasis is on homology and cohomology theory. Throughout this book the only homology groups considered are singular homology groups with integer coefficients, defined by using cubes rather than simplexes. The only cohomology groups considered are the de Rham cohomology groups of a differentiable manifold, defined using C^∞ differential forms. There are a few minor exceptions to these statements which we will note later on. The relative homology and cohomology groups of a pair (X, A) consisting of a topological space X and a subspace A are not considered. As a result, the exact homology or cohomology sequence of such a pair is never used. Instead there is much use of the Mayer-Vietoris exact sequence for a space or manifold which is the union of two open subsets. The homomorphism induced on homology groups by a continuous map is defined, and proved to be invariant under homotopies. The homomorphism induced an cohomology groups by a differentiable map of one manifold to another is also defined.

The reader is introduced to these ideas about homology and cohomology very gently and gradually. First the homology and cohomology groups in dimensions 0

and 1 are defined for an open subset of the plane and applied to various problems. Then it is pointed out that the definitions of the 0- and 1-dimensional homology groups apply without change to an arbitrary topological space. After the introduction of differentiable surfaces (i.e. 2-dimensional manifolds), 2-dimensional cohomology groups are defined (but not 2-dimensional homology groups). Finally in the last part of the book, the homology groups in all dimensions are introduced for arbitrary topological spaces, and cohomology groups in all dimensions are defined for arbitrary differentiable manifolds.

The first ten chapters of the book (150 pages) are concerned mainly with the topology of subsets of the plane. The book commences with a discussion of path integrals in the plane, essentially a review of a topic in advanced calculus. This is followed by the definition of the winding number of a closed path about a point. Several applications of winding numbers are given: a proof of the fundamental theorem of algebra, the Brouwer fixed point theorem for a disc, the Borsuk-Ulam theorem for antipode preserving self maps of the circle and the ham sandwich theorem. Next comes the definition of the 0- and 1-dimensional de Rham cohomology groups and the construction of part of the Mayer-Vietoris exact sequence for two open subsets of the plane. This machinery is sufficient to prove the Jordan Curve theorem in full generality; Brouwer's theorem on "Invariance of Domain" for open subsets of the plane is an easy corollary. These theorems should give the student some idea of the power of these ideas. Other topics in these chapters are the definitions of the 0- and 1-dimensional homology groups and the construction of the Mayer-Vietoris exact homology sequence in dimensions 0 and 1, the definition of the index of a singularity of a vector field in the plane or on a surface, and various classical theorems about such indices of singularities.

The middle part of the book has eight chapters that are devoted mainly to the fundamental group, covering spaces, and the topology of surfaces. The discussion of the fundamental group and the theory of covering spaces is standard, as is the proof of the classification theorem for compact, orientable surfaces (the case of non-orientable surfaces is relegated to an exercise). The fundamental group and first homology group of any surface are determined. Also included is a discussion of the de Rham cohomology groups $H^0(X)$, $H^1(X)$, and $H^2(X)$ of a (differentiable) surface X, the "cup product" pairing $H^1(X) \times H^1(X) \to H^2(X)$ defined by the product of differential forms, and the Mayer-Vietoris exact cohomology sequence for two open subsets of such a surface.

The next part of the book is perhaps its most unusual feature: it consists of three chapters (54 pages) devoted to Riemann surfaces. After the necessary basic definitions, lemmas, etc., these chapters culminate in the statement and proof of the famous Riemann-Roch theorem. Actually this theorem is only proved for Riemann surfaces which arise from a plane algebraic curve, thus avoiding the necessity of proving the existence of a non-constant meromorphic function on an arbitrary Riemann surface. There is also a proof of the Abel-Jacobi theorem which relates the possible integrals on a Riemann surface to its topology. Usually the Riemann-Roch theorem only comes up in a specialized graduate course on algebraic curves or Riemann surfaces. Most universities rarely offer such courses.

The last section, entitled "Higher Dimensions," consists of three chapters (50 pages). Here the homology and cohomology groups are defined in all dimensions and the Mayer-Vietoris exact sequences for homology and cohomology are constructed in full generality. For an arbitrary differentiable manifold M^n the classical de Rham theorem is proved in the following form: integration of q-forms over

differentiable singular q-chains defines a natural homomorphism

$$H^q(M^n) \to Hom[H_q(M^n), \mathbf{R}]$$

which is an isomorphism. From the author's point of view, this famous theorem is the basic relation between homology and cohomology. The degree of a continuous map of an n-sphere to itself is defined, and various standard theorems are proved as a consequence. In another section the Jordan-Brouwer separation theorem about an $(n - 1)$-dimensional sphere in $\mathbf{R}^n$ is proved, along with some corollaries.

In the last chapter the author introduces de Rham cohomology with compact supports by using differential forms with compact supports. The q-dimensional cohomology group of M with compact supports is denoted by $H_c^q(M)$. One may then define a multiplication or pairing $H^k(M) \times H_c^{n-k}(M) \to \mathbf{R}$ for an orientable n-manifold M as follows: take the product of a representative closed k-form and a representative closed $(n - k)$ form with compact support. Integrate this closed n-form over the entire manifold; the result is a real number, which is independent of the choices of representatives. This multiplication defines a homomorphism

$$H^k(M) \to Hom[H_c^{n-k}(M), \mathbf{R}]$$

It is now proved that this homomorphism is an isomorphism for any n-dimensional orientable differentiable manifold. This result is a special case of the general Poincare duality theorem. Tacked on at the end of this last chapter is a section on simplicial complexes. It is proved that the usual homology groups of a finite simplicial complex are isomorphic to the singular homology groups of the underlying topological space. This standard result is not made use of anywhere in the book.

There are several appendices at the end of the book. The last of these appendices is devoted to a proof of the general Borsuk-Ulam theorem about antipodal maps of a sphere to a sphere. This proof uses homology groups with the integers mod 2 as coefficients. It is the only place in the book that homology groups with other than integer coefficients are used.

This book is noteworthy because a lot of motivational material is included to help the student understand the various concepts that are introduced. Probably most textbooks on algebraic topology do not have enough motivational material. To the student, some of these books must seem to consist mostly of algebraic formalisms. Fulton has chosen to develop the subject by means of examples and applications that were historically important, although of course he has not tried to recapitulate the historical development of the subject.

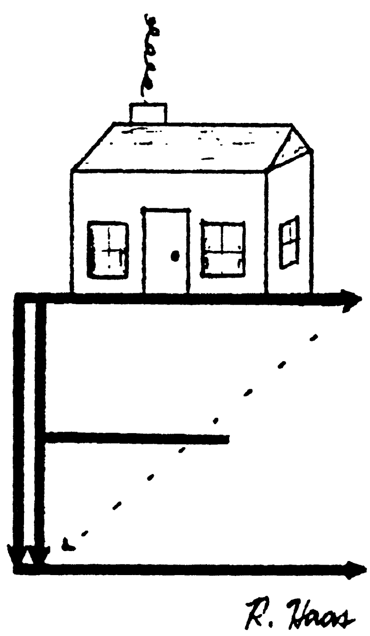
On the other hand, some teachers who are about to give a course on algebraic topology may decide that the book moves too slowly and does not get to some of the important ideas of algebraic topology quickly enough, or perhaps does not get to some of them at all. Unfortunately students nowadays are under great pressure to assimilate a lot of material in a short time, and can not always afford the luxury of a leisurely treatment of a subject, with all the examples and applications it would be desirable to include. This applies not only to graduate students who are struggling to get a Ph.D. in Mathematics; for example, many theoretical physicists nowadays feel it is necessary to learn some of the more subtle ideas of algebraic topology, but have only limited time available.

In this book, as in most books on this subject, many of the motivational examples involve low dimensional topological spaces (i.e., 1, 2, or 3 dimensional). This is in accord with the historical development of algebraic topology. Most of the

definitions and techniques of homology and cohomology theory, which apply to all dimensions, are based on low dimensional examples where geometric intuition is able to guide us. In fact, one of the main tasks of algebraic topology is to deal rigorously with various problems in higher dimensions which are often more or less trivial in low dimensions. An amazing result of this exploration of higher dimensional topology is that often totally new and unexpected phenomena are discovered which have no analogue in lower dimensions. For example, for most values of $n > 6$, the n-dimensional sphere admits several different (i.e. non-isomorphic) structures as a differentiable manifold, while for $n = 1, 2$, or 3 there is only one possibility for a differentiable manifold structure on the n-dimensional sphere (up to isomorphism). The case $n = 4$ is still unsolved. Unfortunately it is not possible to explain exciting examples like this in a beginning text, such as the one under review. Too much machinery is required.

Fulton has done genuine service for the mathematical community by writing a text on algebraic topology which is genuinely different from the existing texts. Each time a text such as this is published we more truly have a real choice when we pick a book for a course or for self-study. The author, who is an expert in algebraic geometry, has given us his own personal idiosyncratic vision of how the subject should be developed.

Department of Mathematics
Yale University



"Homotopy"

Contributed by Robert Haas, Cleveland Heights, Ohio.

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General, P. *Mathematics in St. Petersburg.* A.A. Bolibruch, A.S. Merkur'ev, N. Yu. Netsvetaev. AMS Transl., Ser. 2, V. 174: Adv. in Math. Sci.—30. AMS, 1996, xiii + 273 pp, $99. [ISBN 0-8218-0559-2] A collection of papers by well-known mathematicians who are graduates of the Academic Gymnasium of St. Petersburg University (also known as the "45th Boarding School"). The papers honor the 30th anniversary of the school's founding. AO

Recreational Mathematics, S. *More Rapid Math Tricks and Tips: 30 Days to Number Mastery.* Edward H. Julius. Wiley, 1996, xi + 226 pp, $12.95 (P). [ISBN 0-471-12238-6] Several dozen specialized methods for performing mental arithmetic (e.g., squaring numbers ending in 6). Intended for parlor tricks, sales negotiations, and failed-battery moments. Eerily reminiscent of ancient arithmetic texts in which each problem entails a different algorithm. LAS

Education, P, L. *Aptitude Revisited: Rethinking Math and Science Education for America's Next Century.* David E. Drew. Johns Hopkins Univ Pr, 1996, xii + 254 pp, $35.95. [ISBN 0-8018-5143-2] A well-documented review of experience and research on who gets to learn mathematics and science in the U.S. today, and on our dysfunctional educational system "through which the inequalities of our society are perpetrated and exacerbated." Includes extensive discussion of Treisman's Berkeley calculus "workshop groups" and Tobias' observations of student experiences. A useful reference that synthesizes and documents a quarter-century's evidence of systemic failure to educate all students. LAS

History, P. *History of Mathematics: States of the Art.* Eds: Joseph W. Dauben, *et al.* Academic Pr, 1996, xxiv + 394 pp, $59.95. [ISBN 0-12-204055-4] Narrowly focused articles on a broad range of subjects, Thabit ibn Qurra to Abraham Robinson. Most in German. DB

History, L. *Through a Reporter's Eyes: The Life of Stefan Banach.* Roman Kałuża. Transl. & Ed: Ann Kostant, Wojbor Woyczyński. Birkhäuser Boston, 1996, x + 137 pp, $24.50. [ISBN 0-8176-3772-9] For non-mathematicians. Describes the Polish mathematical community of the first half of this century. DB

History, S(17), L. *Italian Algebraic Geometry between the Two World Wars.* A. Brigaglia, C. Ciliberto. Papers in Pure & Appl. Math., V. 100. Queen's Univ, 1995, viii + 223 pp, (P). [ISBN 0-88911-699-7] Interesting collection of essays give broad historical background. Translated from the Italian. RM

Logic, T(17: 1), P. *Extensions of First Order Logic.* María Manzano. Tracts in Theoret. Comp. Sci. Cambridge Univ Pr, 1996, xxii + 388 pp, $59.95. [ISBN 0-521-35435-8] Many sorted logics (with sets of variables—and quantification) over several universes of objects provide a generalization of first-order logic which the author elaborates as a unifying theme for second order, modal, and dynamic logics, type theory, etc. Examples and motivating discussions connect the theory with applications to

foundations of mathematics, computer science, and linguistics. RM

Logic, P. *Non-Standard Analysis, Revised Edition.* Abraham Robinson. Landmarks in Math. Princeton Univ Pr, 1996, xix + 293 pp, $19.95 (P). [ISBN 0-691-04490-2] Reprint of the 1974 revised edition.

Logic, P. *Computability, Enumerability, Unsolvability: Directions in Recursion Theory.* Eds: S.B. Cooper, T.A. Slaman, S.S. Wainer. London Math. Soc. Lect. Note Ser., V. 224. Cambridge Univ Pr, 1996, vii + 347 pp, $39.95 (P). [ISBN 0-521-55736-4] 15 research papers commemorating the 1993–94 Leeds Recursion Theory Year.

Combinatorics, T(17), P. *Combinatorial Methods in Discrete Mathematics.* Vladimir N. Sachkov. Ency. of Math. & Its Applic., V. 55. Cambridge Univ Pr, 1996, xiii + 306 pp, $69.96. [ISBN 0-521-45513-8] Topics on existence and construction (Latin squares, block designs, transversals, permanents, etc.) and enumeration (generating functions, Pólya's theorem, etc.). Central theme is the general combinatorial scheme based on the notion of a mapping. Also includes discussion of asymptotics (saddle point method). LC

Discrete Mathematics, T*(15–17), L. *An Introduction to Difference Equations.* Saber N. Elaydi. Undergrad. Texts in Math. Springer-Verlag, 1996, xiii + 389 pp, $45. [ISBN 0-387-94582-2] Well-written. Features an extensive collection of applications. Emphasizes stability and control theory; omits combinatorics, number theory, chaos theory. Exercises encourage calculator/computer exploration. TH

Discrete Mathematics, P, L*. $A = B$. Marko Petkovšek, Herbert S. Wilf, Doron Zeilberger. AK Peters, 1996, xii + 212 pp, $39. [ISBN 1-56881-063-6] The mathematics behind algorithms and methods for finding hypergeometric sums. Includes software for *Maple* and *Mathematica*. LC

Number Theory, P. *Limit Theorems for the Riemann Zeta-Function.* Antanas Laurinčikas. Math. & Its Applic., V. 352. Kluwer Academic, 1996, xiii + 297 pp, $149. [ISBN 0-7923-3824-3] Probability theory used to study distribution of values of Dirichlet series. DB

Number Theory, P, L.** *Notes on Fermat's Last Theorem.* Alf van der Poorten. Canadian Math. Soc. Ser. of Mono. & Adv. Texts. Wiley, 1996, xvi + 222 pp, $44.95. [ISBN 0-471-06261-8] Seventeen informal "lectures" that revolve about Wiles' proof of Fermat's Last Theorem. Purpose is "to glimpse all sorts of exciting pieces of mathematics and to be moved to teach ourselves more." Extensive references. A must for every undergraduate library and for anyone who wants an accessible introduction to the ideas that have played a role in the pursuit of this theorem. DB

Number Theory, T(13: 1), S*. *A Cascade of Numbers: An Introduction to Number Theory.* Bob Burn, Amanda Chetwynd. Edward Arnold, 1996, ix + 148 pp, (P). [ISBN 0-340-65251-9] Nicely motivated, problem- and puzzle-oriented introduction to number theory. Each of 53 sections has initial explore and discover part; complement gives comments, solutions, proofs, and theoretical development. Woven throughout are a variety of strands: primes, congruences, Euler's ϕ, residues, Fermat, proof techniques. RM

Group Theory, T(16–18: 1), P, L. *Permutation Groups.* John D. Dixon, Brian Mortimer. Grad. Texts in Math., V. 163. Springer-Verlag, 1996, xii + 346 pp, $49. [ISBN 0-387-94599-7] Largely self-contained. Includes chapters on primitive groups with a proof of the O'Nan–Scott theorem, Mathieu groups and Steiner systems, multiply transitive groups, infinite permutation groups. Appendices on classification of finite simple groups, and a listing of primitive permutation groups of degree less than 1000. JS

Group Theory, S(16–18), P, L. *Symmetry Orbits.* Hugo F. Verheyen. Design Sci. Coll. Birkhäuser Boston, 1996, vii + 236 pp, $75. [ISBN 0-8176-3661-7] Aimed at a broad audience. Part I includes an introduction to symmetry groups, an extensive catalog of groups, and general discussion of orbits under symmetry groups. Part II deals with compounds of cubes obtained by having groups act on a cube. Beautifully illustrated with diagrams and photos of models. Includes a chapter on making models. JS

Algebra, P. *Lie Algebras and Their Representations.* Eds: Seok-Jin Kang, Myung-Hwan Kim, Insok Lee. Contemp. Math., V. 194. AMS, 1996, viii + 232 pp, $45 (P). [ISBN 0-8218-0512-6] Proceedings of a 1995 symposium at Seoul National University.

Algebra, T(18: 1), S, P. *Rings with Generalized Identities.* K.I. Beidar, W.S. Martindale, III, A.V. Mikhalev. Pure & Appl. Math., V. 196. Marcel Dekker, 1996, xi + 522 pp, $185. [ISBN 0-8247-9325-0] A thorough and largely self-contained treatment, including recent work of Kharchenko and Chuang. First four chapters are foundation for later application to Lie algebras (PBW theorems), T-identities for prime

and semiprime rings, and Lie theory including resolution of a conjecture of Herstein. JS

Algebra, P. *Algebra.* Eds: Yuriĭ, L. Ershov, *et al.* Walter de Gruyter, 1996, xiii + 304 pp, DM 268. [ISBN 3-11-014413-1] Proceedings of a 1993 conference in Krasnoyarsk, Russia.

Algebra, P. *Representation Theory of Algebras and Related Topics.* Eds: Raymundo Bautista, Roberto Martínez-Villa, José Antonio de la Peña. Canadian Math. Soc. Conf. Proc., V. 19. AMS, 1996, xvii + 406 pp, $95 (P). [ISBN 0-8218-0396-4] Proceedings of a 1994 workshop at UNAM, Mexico City.

Algebra, P. *Representation Theory of Algebras.* Eds: Raymundo Bautista, Roberto Martínez-Villa, José Antonio de la Peña. Canadian Math. Soc. Conf. Proc., V. 18. AMS, 1996, xxi + 749 pp, $129 (P). [ISBN 0-8218-0395-6] Proceedings of a 1994 conference in Cocoyoc, Mexico.

Algebra, P. *An Introduction to Nonassociative Algebras.* Richard D. Schafer. Dover, 1995, x + 166 pp, $6.95 (P). [ISBN 0-486-68813-5] Republication, with corrections, of the 1966 Academic Press edition (TR, August–September 1967).

Algebra, P. *Abstract Algebra.* W.E. Deskins. Dover, 1995, xiii + 624 pp, $15.95 (P). [ISBN 0-486-68888-7] Republication, with corrections, of the second (1966) printing of the 1964 Macmillan edition.

Calculus, T(13–14: 1). *Brief Calculus: A Graphing Calculator Approach.* Ruric Wheeler, Karla Neal, Roseanne Hofmann. Wiley, 1996, xv + 540 pp, $70.95. [ISBN 0-471-05721-5] A standard brief calculus with a graphing calculator approach incorporating many nice problems. Worth considering. PF

Calculus, T(13–14: 1). *Calculus Lite.* Frank Morgan. AK Peters, 1995, xiv + 281 pp, $29.95. [ISBN 1-56881-037-7] A nice, concise introduction to differential and integral calculus with a nod towards infinite series and differential equations. PF

Differential Equations, S. *Schaum's Outline of Theory and Problems of Differential Equations, Second Edition.* Richard Bronson. McGraw-Hill, 1994, x + 358 pp, $13.95 (P). [ISBN 0-07-008019-4] Changes from *First Edition* include a new section on direction fields, some reorganization, and many new problems. Over 1300 problems in total. SK

Differential Equations, T(17: 1), P. *Introduction to Linear Systems of Differential Equations.* L. Ya. Adrianova. Transl. of Math. Mono., V. 146. AMS, 1995, x + 204 pp, $99. [ISBN 0-8218-0328-X] Basic theory of autonomous, periodic systems: reducibility, regularity, stability, and solution growth rates. Includes results previously available only in specialized journals. SK

Differential Equations, T(14–15: 1). *Ordinary Differential Equations.* W. Cox. Modular Math. Ser. Edward Arnold, 1996, x + 222 pp, (P). [ISBN 0-340-63203-8] Readable, superficial treatment of standard topics. SK

Differential Equations, T(14), S*, L*. *Differential Equations: A Dynamical Systems Approach: Higher-Dimensional Systems.* John H. Hubbard, Beverly H. West. Texts in Appl. Math., V. 18. Springer-Verlag, 1995, xiv + 601 pp, $49 (P). [ISBN 0-387-94377-3] Volume two of ambitious, terrific four-volume series. Very geometric approach, emphasis on "understanding the behavior of differential equations, rather than on solving" (from the Preface). Deep treatment of autonomous, non-linear, planar systems including chapters on structural stability and bifurcations. Really nice exercises. Supported by *MacMath* software package (not included). Give it a look before teaching ODE's again. SK

Partial Differential Equations, T(18: 1), P. *Exponential Attractors for Dissipative Evolution Equations.* A. Eden, *et al.* Res. in Appl. Math. Wiley, 1994, viii + 182 pp, $39.95 (P). [ISBN 0-471-95223-0] Sometimes dissipative PDE's have finite dimensional global attractors, sometimes it is possible to realize these attractors as solutions to an ODE. This greatly simplifies analysis of the dynamics. Of course, sometimes the latter is not possible and then perhaps it is possible to construct an *exponential attractor*, an exponentially attracting, finite dimensional, forward flow-invariant, compact set which contains the global attractor. Construction, approximation, and examples of exponential attractors. SK

Partial Differential Equations, P. *Geometric and Algebraic Structures in Differential Equations.* Eds: P.H.M. Kersten, I.S. Krasil'shchik. Kluwer Academic, 1995, vi + 348 pp, $168. [ISBN 0-7923-3871-5] Invited lectures from a 1993 workshop at the University of Twente. Reprinted from *Acta Applicandae Mathematicae*, V. 41, Nos. 1–3 (1995).

Partial Differential Equations, P. *Partial Differential Equations and Mathematical Physics.* Eds: Lars Hörmander, Anders Melin. Prog. in Nonlinear Diff. Eqns. & Their Applic., V. 21. Birkhäuser Boston, 1996, 372 pp, $89.50. [ISBN 0-8176-3906-3] Papers from seminars

given at the Universities of Copenhagen and Lund in 1995.

Partial Differential Equations, P. *Mathematical Modelling of Heat and Mass Transfer Processes.* V.G. Danilov, V.P. Maslov, K.A. Volosov. Math. & Its Applic., V. 348. Kluwer Academic, 1995, ix + 316 pp, $159. [ISBN 0-7923-3789-1]

Dynamical Systems, P. *The Cauchy Problem in Kinetic Theory.* Robert T. Glassey. SIAM, 1996, xii + 241 pp, $49 (P). [ISBN 0-89871-367-6] A thorough treatment of what is known about initial value problems for the Boltzmann equation for rarefied gas dynamics, and the Vlasov–Poisson and Vlasov–Maxwell systems from plasma physics. DB

Numerical Analysis, P. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* J.E. Dennis, Jr., Robert B. Schnabel. SIAM, 1996, xv + 378 pp, $34.50 (P). [ISBN 0-89871-364-1] Republication, with corrections, of the 1983 Prentice Hall edition (TR, January 1984).

Numerical Analysis, P. *Linear and Nonlinear Conjugate Gradient-Related Methods.* Eds: Loyce Adams, J.L. Nazareth. SIAM, 1996, xvi + 164 pp, $32 (P). [ISBN 0-89871-376-5] Proceedings of a 1995 conference at the University of Washington.

Numerical Analysis, P. *Applications of Interval Computations.* Eds: R. Baker Kearfott, Vladik Kreinovich. Appl. Optim., V. 3. Kluwer Academic, 1996, xvii + 425 pp, $182. [ISBN 0-7923-3847-2] Papers from a 1995 workshop in El Paso.

Operator Theory, P. *Characteristic Functions and Models of Nonself-Adjoint Operators.* A. Kuzhel. Math. & Its Applic., V. 349. Kluwer Academic, 1996, ix + 273 pp, $159. [ISBN 0-7923-3879-0]

Operator Theory, T(18), P. *A User's Guide to Operator Algebras.* Peter A. Fillmore. Canadian Math. Soc. Ser. of Mono. & Adv. Texts. Wiley, 1996, xii + 223 pp, $59.95. [ISBN 0-471-31135-9] Deciphers the recent Fields Medal works of Alain Connes (1982) and Vaughan Jones (1990) for the non-specialist "as painlessly as possible." Background in operator theory required. First five chapters provide a comprehensive introduction to C^*-algebras and von Neumann algebras. SA

Operator Theory, T(18), S, P. *Linear Operators and Ill-Posed Problems.* M.M. Lavrent'ev, L. Ya. Savel'ev. Transl: Nauka Publishers, Moscow. Consultants Bureau, 1995, xii + 382 pp, $110. [ISBN 0-306-11035-0] A well-posed initial-value problem is one where the solution (i) exists, (ii) is unique, and (iii) depends continuously on the data. Ill-posed refers mainly to (iii). Without it, a solution is actually not determined since in a real physical setting the data are only approximations. First three chapters cover the required operator theory. SA

Operator Theory, S(18), P. *Spectral Theory of Differential Operators: Self-Adjoint Differential Operators.* V.A. Il'in. Consultants Bureau, 1995, xvii + 390 pp, $115. [ISBN 0-306-11037-7] A systematized summary of the major results obtained by the author in the field of the spectral theory of self-adjoint elliptic operators. Contains many previously unreported results. Techniques are primarily based on the mean-value theorems for regular solutions of elliptic equations with a spectral parameter. SA

Functional Analysis, P. *Realizability Theory for Continuous Linear Systems.* A.H. Zemanian. Dover, 1995, xv + 231 pp, $9.95 (P). [ISBN 0-486-68823-2] Republication, with corrections, of the 1972 Academic Press edition (TR, April 1973).

Functional Analysis, P. *Classical Banach Spaces I–II.* Joram Lindenstrauss, Lior Tzafriri. Classics in Math. Springer-Verlag, 1996, $35 (P), [ISBN 3-540-60628-9]. *I: Sequence Spaces*, xiii + 188 pp; *II: Function Spaces*, x + 243 pp. Republication in a single volume of works originally published in 1977 and 1979 (TR, March 1978 and February 1980).

Functional Analysis, P. *Vector Lattices and Integral Operators.* Ed: S.S. Kutateladze. Math. & Its Applic., V. 358. Kluwer Academic, 1996, ix + 462 pp, $235. [ISBN 0-7923-3897-9] An introduction to the theory of vector lattices. A revised and updated translation of the original 1992 Russian edition.

Functional Analysis, T(17). *Fundamentals of Functional Analysis.* S.S. Kutateladze. Texts in Math. Sci., V. 12. Kluwer Academic, 1996, xiv + 276 pp, $170. [ISBN 0-7923-3898-7] Concise introduction to the essential concepts of modern functional analysis. Discusses Banach and Hilbert spaces, the Riesz–Dunford functional calculus, Fredholm theory, and duality theory for locally convex spaces. Includes a final chapter on Banach algebras. A solid text and useful reference for all the famous theorems of the subject. SA

Functional Analysis, P. *Functional Analysis on the Eve of the 21st Century, Volumes I–II.* Eds: Simon Gindikin, James Lepowsky, Robert L. Wilson. Birkhäuser Boston, 1995, $84.50 set, [ISBN 0-8176-3860-1] set. *Volume I*, Prog.

in Math., V. 131, xxii + 292 pp; *Volume II*, Prog. in Math., V. 132, xxii + 323 pp. Papers from a 1993 conference at Rutgers University honoring the eightieth birthday of I.M. Gelfand.

Analysis, P. *World Congress of Nonlinear Analysts '92, Volumes I–IV.* Ed: V. Lakshmikantham. Walter de Gruyter, 1996, DM 1.198 set, [ISBN 3-11-013215-X] set. *Volume I*, xlvi + 1016 pp; *Volume II*, xi + 1011 pp; *Volume III*, xi + 1013 pp; *Volume IV*, x + 911 pp. Proceedings of the 1992 conference in Tampa, Florida.

Algebraic Geometry, P. *Collected Papers of Ruggiero Torelli.* Eds: Ciro Ciliberto, Paulo Ribenboim, Edoardo Sernesi. Papers in Pure & Appl. Math., V. 101. Queen's Univ, 1995, xii + 224 pp, (P). [ISBN 0-88911-707-1]

Algebraic Geometry, P. *Current Topics in Complex Algebraic Geometry.* Eds: Herbert Clemens, János Kollár. Math. Sci. Res. Inst., V. 28. Cambridge Univ Pr, 1995, xi + 158 pp, $34.95. [ISBN 0-521-56244-9] 8 survey articles from the 1992–93 academic year at MSRI.

Algebraic Geometry, P. *Topology of Real Algebraic Varieties and Related Topics.* Eds: V. Kharlamov, *et al.* AMS Transl., Ser. 2, V. 173. AMS, 1996, vi + 264 pp, $99. [ISBN 0-8218-0555-X] Papers written by friends, students, and collaborators of D.A. Gudkov.

Algebraic Geometry, P. *The Curves Seminar at Queen's, Volume X.* Anthony V. Geramita. Papers in Pure & Appl. Math., No. 102. Queen's Univ, 1996, 363 pp, (P). [ISBN 0-88911-717-9]

Differential Geometry, P. *New Developments in Differential Geometry.* Eds: L. Tamássy, J. Szenthe. Math. & Its Applic., B. 350. Kluwer Academic, 1996, x + 434 pp, $209. [ISBN 0-7923-3822-7] Proceedings of a 1994 conference in Debrecen, Hungary.

Topology, P. *Novikov Conjectures, Index Theorems and Rigidity, Volumes 1–2.* Eds: Steven C. Ferry, Andrew Ranicki, Jonathan Rosenberg. Cambridge Univ Pr, 1995, $39.95 (P) each. *Volume 1*, London Math. Soc. Lect. Note Ser., V. 226, x + 372 pp, [ISBN 0-521-49796-5]; *Volume 2*, London Math. Soc. Lect. Note Ser., V. 227, x + 364 pp. [ISBN 0-521-49795-7] Includes a detailed historical survey; an annotated problems list; the texts of some important unpublished papers; several research and survey papers.

Topology, P. *Global Surgery Formula for the Casson–Walker Invariant.* Christine Lescop. Annals of Math. Stud., No. 140. Princeton Univ Pr, 1996, 150 pp, $49.50 (P). [ISBN 0-691-02132-5]

Operations Research, S(17), P. *Control of Freeway Traffic Flow.* S.A. Smulders. CWI Tract, V. 80. Centrum voor Wiskunde en Informatica, 1996, 158 pp, Dfl. 40 (P). [ISBN 90-6196-451-2] Nicely developed model and analysis of traffic flow, including density estimation, filtering, stability, control, and simulations. RM

Optimization, P. *Minimax and Applications.* Eds: Ding-Zhu Du, Panos M. Pardalos. Nonconvex Optim. & Its Applic., V. 5. Kluwer Academic, 1995, xiv + 293 pp, $149. [ISBN 0-7923-3615-1] 18 papers survey recent research.

Probability, T(17: 1), L. *Probability Theory.* Heinz Bauer. Transl: Robert B. Burckel. Stud. in Math., V. 23. Walter de Gruyter, 1996, xv + 523 pp, DM 148. [ISBN 3-11-013935-9] Introduction to probability theory and stochastic processes. Assumes knowledge of measure and integration theory. Topics: basic theory, independence, law of large numbers, martingales, Fourier analysis, limit distributions, law of the iterated logarithm, construction of stochastic processes, Brownian motion. KB

Probability, S(17), P, L. *The Exponential Distribution: Theory, Methods and Applications.* Eds: N. Balakrishnan, Asit P. Basu. Gordon & Breach, 1995, xxv + 638 pp, $119. [ISBN 2-88449-192-9] A systematic and comprehensive synthesis of the literature on the exponential distribution. KB

Stochastic Processes, P. *Diffusion Processes and their Sample Paths, Second Printing, Corrected.* Kiyosi Itô, Henry P. McKean, Jr. Classics in Math. Springer-Verlag, 1996, xv + 321 pp, $35 (P). [ISBN 3-540-60629-7] Reprint of the 1974 edition (TR, April 1975).

Elementary Statistics, T*(13: 1), S. *Activity-Based Statistics.* Richard L. Scheaffer, *et al.* Springer-Verlag, 1996. *Instructor Resources*, xii + 359 pp, $42.95, [ISBN 0-387-94597-0]; *Student Guide*, xi + 252 pp, $24, (P). [ISBN 0-387-94598-9] A wealth of engaging, innovative student activities for most major topics. Provides a framework for exploring the relationship between data collection, data analysis, and problem solving. Stresses the development of a conceptual understanding of sampling error, confidence intervals, and the p-value. Notes for instructor include very helpful sample results. Highly recommended. RS

Elementary Statistics, T(13: 1). *Engineering Statistics: The Industrial Experience.* Bernard Ostle, *et al.* Duxbury Pr, 1996, xiv + 609 pp, $64.50. [ISBN 0-534-26538-3] Topics: de-

scriptive statistics, graphical displays, probability functions and distributions, sampling distributions and estimation, tests of hypotheses, Shewhart control charts, measurement system evaluation, process capability studies, regression analysis, design of experiments. Uses real industrial data. Not dependent on a particular computer package. KB

Mathematical Statistics, T(16–17: 1, 2). *Mathematical Statistics for Economics and Business.* Ron C. Mittelhammer. Springer-Verlag, 1996, xviii + 723 pp, $54.95. [ISBN 0-387-94587-3] Standard topics in introductory mathematical statistics. Numerous applications illustrate the theory. Exercises ask students to apply theory and to do proofs (emphasis heavily on the former). Extensive inclusion of applications accomplished by increasing length, rather than reducing treatment of theory. Requires multivariable calculus, matrix arithmetic. LB

Statistical Methods, T(17–18), P. *A Guide to Chi-Squared Testing.* Priscilla E. Greenwood, Mikhail S. Nikulin. Ser. in Prob. & Stat. Wiley, 1996, xii + 280 pp, $39.95. [ISBN 0-471-55779-X] Issues and recent developments in chi-square testing. Includes several measures of fit, issues involved in testing after parameter estimation, grouping data, and best unbiased estimators for an exponential family of distributions. RS

Statistical Methods, T(18), P. *Robust Statistical Procedures: Asymptotics and Interrelations.* Jana Jurečková, Pranab Kumar Sen. Ser. in Prob. & Stat. Wiley, 1996, xiv + 466 pp, $59.95. [ISBN 0-471-82221-3] Unified treatment of the asymptotics of robust statistics. Part I discusses theoretical foundations including first- and second-order asymptotic distributional representations for M, L, and R estimators. Part II discusses statistical inference based on these robust statistics/estimators. RS

Statistics, P. *Stochastic Modelling in Physical Oceanography.* Eds: Robert J. Adler, Peter Müller, Boris Rozovskii. Prog. in Prob., V. 39. Birkhäuser Boston, 1996, xi + 467 pp, $84.50. [ISBN 0-8176-3798-2] A collection of expository-survey articles accessible to researchers in either probability and statistics or physical oceanography.

Statistics, P. *Efficient and Inefficient Estimation in Semiparametric Models.* M.J. van der Laan. CWI Tract, V. 114. Centrum voor Wiskunde en Informatica, 1995, vii + 219 pp, Dfl. 50 (P). [ISBN 90-6196-461-X]

Programming, P. *C Programming: A Modern Approach.* K.N. King. WW Norton, 1996, xxv + 661 pp, $39.95 (P). [ISBN 0-393-96945-2]

Computer Systems, P. *World Wide Web Journal.* Chairs: Albert Vezza, Ira Goldstein. O'Reilly Associates, 1995, xii + 735 pp, $39.95 (P). [ISBN 1-56592-169-0] Proceedings of the Fourth International World Wide Web Conference (Boston, December 1995).

Computer Systems, P. *Perl 5: Desktop Reference.* Johan Vromans. O'Reilly & Associates, 1996, v + 39 pp, $6.95 (P). [ISBN 1-56592-187-9]

Computer Systems, C, P. *Migrating to Windows 95: A Programmer's Guide to What's New.* Mark Andrews. Academic Pr, 1996, xxix + 623 pp, $34.95 (P), with disk. [ISBN 0-12-058525-1]

Computer Systems, P. *The X Resource, Issue Seventeen: Proceedings of the 10th Annual X Technical Conference.* Ed: Paula M. Ferguson. O'Reilly & Associates, 1996, 252 pp, $14.95 (P). [ISBN 1-56592-178-X]

Computer Science, S(16–17: 1), P. *Fuzzy Databases: Principles and Applications.* Frederick E. Petry. Kluwer Academic, 1996, xi + 226 pp, $94. [ISBN 0-7923-9667-7] Accessible development of recent research into the theory and commercial application of fuzzy set theory to extend relational (and object-oriented) databases to better deal with the representation and manipulation of uncertain information. Application involving Graphical Information System (GIS) is nice motivating example, developed throughout. RM

Computer Science, P. *Logic Programming.* Ed: John Lloyd. MIT Pr, 1995, xx + 644 pp, $75 (P). [ISBN 0-262-62099-5] Proceedings of the 1995 International Symposium in Portland, Oregon.

Computer Science, P. *Natural & Artificial Parallel Computation: Proceedings of the Fifth NEC Research Symposium.* Ed: David L. Waltz. SIAM, 1996, vii + 202 pp, $39. [ISBN 0-89871-357-9]

Applications (Communication Theory), P. *Combinatorial Network Theory.* Eds: Ding-Zhu Du, D. Frank Hsu. Appl. Optim., V. 1. Kluwer Academic, 1996, vii + 212 pp, $99. [ISBN 0-7923-3777-8] 5 papers on the interconnection problem.

Reviewers

SA: Steve Abbott, St. Olaf; KB: Karla Ballman, Macalester; LB: Lynne Baur, Carleton; DB: David Bressoud, Macalester; LC: Laura Chihara, St. Olaf; PF: Paul Froeschl, Macalester; TH: Tom Halverson, Macalester; SK: Steve Kennedy, Carleton; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; JS: John Schue, Macalester; RS: Richard Single, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf.


TIAA-CREF
rated #1
in customer
satisfaction.
The participants have spoken. In the
most recent Dalbar survey of
tomer satisfaction. TIAA-CREF
rated highest,
most well-

Lion Hunting

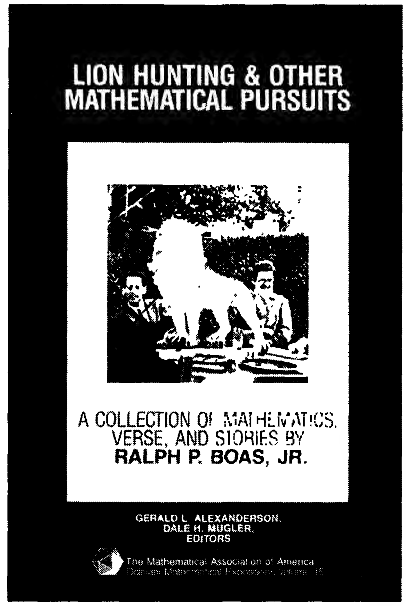
.nd Other Mathematical Pursuits

Collection of Mathematics, Verse, and Stories
r Ralph P. Boas, Jr.

ɜrald L. Alexanderson and
ıle H. Mugler, Editors

ıighly recommend Lion Hunting and Other athematical Pursuits to high school mathematics ıbs, mathematics teachers of all levels, and anyone terested in mathematics. Perhaps the most impor- nt features of this book is how it subtly makes the ader aware of the nature of mathematics.

– The Mathematics Teacher



ꞩ a young man at the Institute for Advanced Study in inceton, Ralph Philip Boas, Jr., together with a group of her mathematicians, published a light-hearted article on e "mathematics of lion hunting" under a pseudonym ꝰ38). This sparked a sequence of articles on the topic, veral of which are drawn together in this book.

on Hunting includes an assortment of articles that show e many facets of this remarkable mathematician, editor, riter, and teacher. Along with a variety of his lighter athematical papers, the collection includes Boas' verse ıd short stories, many of which are appearing for the first ne. Anecdotes and recollections of his numerous experi- ıces and of his work and meetings with many distin- ıished mathematicians and scientists of his day are also cluded as well as photographs taken by Boas of Hardy, ttlewood, Besicovitch, Weil, and others.

ıe mathematical articles in this collection cover a range topics. They include articles on infinite series, the mean ılue theorem, indeterminate forms, complex variables, verse functions, extremal problems for polynomials and ore. A special section of this book is devoted to articles ›out the teaching of mathematics, with titles such as "Calculus as an experimental science" and "Can we mak mathematics intelligible?"

Boas's wit and playful humor are reflected in the versє included in this collection. The verses reflect the phases (his career as author, editor, teacher, department chair, an lover of literature. A section of the book describes the feu that Boas supposedly had with Bourbaki. Also included aı many amusing anecdotes about famous mathematicians.

320 pp., Paperbound, 1995, ISBN 0-88385-323-X
List: $39.95 MAA Member: $28.95
Catalog Code: DOL-15

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
P.O. Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

ſembership Code: ___ ___ ___ ___ ___ ___

ſame __________

ddress __________

꞉ity __________

tate __________ Zip __________

QTY.	CATALOG CODE	PRICE	AMOUN
____	DOL-15	____	____
____	____	____	____
		TOTAL	____

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. __________ Expires __/__

Signature __________

INDEX TO VOLUME 103, 1996
THE AMERICAN MATHEMATICAL MONTHLY

TITLE INDEX

AUTHOR INDEX

PROBLEMS PROPOSED

PROBLEMS SOLVED

SOLUTIONS

Numbers in boldface refer to problems; those in lightface to pages.

THANKS

The Monthly expresses its appreciation to the following people for their help in refereeing during the past year. We could not function successfully without such people and their hard work.

Janós Aczel
John Akeroyd
George Baloglou
Jozsef Beck
Bruce C. Berndt
Jonathan M. Borwein
Nigel Boston
Bart Braden
Tom C. Brown
Richard A. Brualdi
David Cantor
Gulbank D. Chakerian
Gregory L. Cherlin
John H. Conway
Dennis DeTurck
John D. Dixon
Clayton W. Dodge
John Duncan
Peter Duren
H. Farahat
Fred Galvin
George Gasper
Walter Gautschi
George T. Gilbert
Bart E. Goddard
Chaim Goodman-Strauss
Hardy Grant
Andrew J. Granville
Alfred Gray
H. W. Guggenheimer
Richard K. Guy
Michael Hurley
I. Marty Isaacs
Charles Johnson
Mark Johnson
Geoffrey A. Kandall
Jon Kane
Clark Kimberling
Victor Klee
Mark I. Krusemeyer
Jeffrey C. Lagarias
Peter Lancaster
Richard N. Lyons
Gyula Maksa
Itrel Monroe
Howard Morris
Jan Mycielski
Roger Nussbaum
Bjorn Poonen
Mizanur Rahman
David P. Robbins
James R. Roche
Michael Saks
Jonathan Schaer
Stephen Schanuel
Hans Schneider
Sanford L. Segal
Marjorie Senechal
Lawrence A. Shepp
Joseph H. Silverman
Leonard Smiley
Paulo Starni
Michael J. Steele
Serge Tabachnikov
Patrick Tantalo
Charles Toll
Jerry Uhl
Stan Wagon
Herbert Wilf
Tudor Zamfirescu
Doron Zeilberger
Cunhui Zhang

THE AMERICAN MATHEMATICAL

MONTHLY

Volume 103, Number 10 December 1996

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The *Monthly* publishes articles, notes, and other features about mathematics and the profession. The readership of the *Monthly* is intended to include everybody who is mathematically inclined, including of course professional mathematicians and students of mathematics at all collegiate levels. While no single article or feature is likely to appeal to everyone, material should interest and be accessible to a large number of readers. This is the most important criterion for acceptance.

Articles may be expositions of old results or presentations of new ones. They may concern all of mathematics or one small area, a broad development or a single application, historical reminiscences or one important event. While some articles may contain the author's new research, the novelty of material and generality of the results is far less important than the clarity of exposition and general interest. Discussing one illuminating case of a well known result is far better than providing all the details of an obscure but new proposition. Articles in the *Monthly* are supposed to inform and to entertain; they are meant to be read rather than archived.

Notes are short and possibly informal articles. A note may concern a clever new proof of an old theorem, a novel way to present tired material, or a lively discussion of a philosophical (but still mathematical) issue. Also, any topic is suitable, so long as it is related to mathematics. Because a note is short, the first few sentences are the most important part: They should explain the purpose and invite the reader in. Photographs or diagrams often will attract the reader's attention.

All articles and notes should be sent to the Editor-Elect:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send 3 copies, typewritten on only one side of the paper. Illustrations should be carefully drawn on separate sheets of paper in black ink; the original should be without lettering and two copies should have appropriate captions and lettering indicated.

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all material, typewritten on only one side of the paper.

Letters to the Editor, both for publication and for private reading, should be sent to the Editor-Elect at the address given above. Comments, including criticisms, are welcome, as are all suggestions for making the *Monthly* a lively, entertaining, and informative journal.

Reprint permission:
MARCIA P. SWARD, Executive Director

Advertising Correspondence:
Ms. ELAINE PEDREIRA, Advertising Manager

Subscription correspondence, change of address, and other inquiries:
Membership / Subscriptions Department

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

Second class postage paid at Washington, DC, and additional mailing offices. ***Postmaster***: Send address changes to the American Mathematical Monthly, Membership / Subscription Department, MAA, 1529 Eighteenth Street, N.W., Washington, DC, 20036-1385.

The Mathematical Collaboration of M. L. Cartwright and J. E. Littlewood

Shawnee L. McMurran and James J. Tattersall

Balthasar van der Pol's experiments on electrical circuits during the 1920s and 1930s opened an interesting chapter in the history of dynamics. The need for advancements in radio technology made van der Pol's work pertinent and his research stimulated mathematical interest in nonlinear oscillators. In particular, van der Pol's work caught the attention of Cambridge mathematicians M. L. Cartwright and J. E. Littlewood. Topology and Poincaré's transformation theory provided a key to analyzing behavior of nonlinear oscillators and dissipative systems. Resulting mathematical techniques have played a significant role in the development of the modern theory of dynamical systems and chaos. In addition, nonlinear oscillator theory has led to the development of radio, radar, and laser technology.

The collaboration between Cartwright and Littlewood began just before World War II and lasted approximately ten years. They published four joint papers, and individually published several other papers based on joint work. Their collaboration produced some of the earliest rigorous work in the field of large parameter theory. They were among the first mathematicians to recognize that topological and analytical methods could be combined to efficiently obtain results for various problems in differential equations, and their results helped inspire the construction of Smale's horseshoe diffeomorphism.[1] Their names can be included with those mathematicians, including Levinson, Lefschetz, Minorsky, Liapounov, Kryloff, Bogolieuboff, Denjoy, Birkhoff, and Poincaré, whose work provided an impetus to the development of modern dynamical theory.

Cartwright knew G. H. Hardy at Oxford before she met Littlewood. Cartwright joined a special group when she began to attend Hardy's Friday class at New College in January, 1928. The participants included Gertrude Stanley, John Evelyn, E. H. Linfoot, L. S. Bosenquet, Frederick Brand, and Tirukkannapuram Vijayarhagavan. Nearly all completed their D. Phils by the summer of 1928. The universal feeling among the participants was that they were studying under a very great man who had not yet been recognized fully. Cartwright appreciated Hardy's style and philosophy of mathematics. She recalled that he took immense trouble with his students whether they were good, bad, or indifferent. Once, when she had produced an obviously fallacious result, Hardy remarked, "Let's see, there's always hope when you get a sharp contradiction."[2] Hardy became Cartwright's initial thesis advisor and she finished with E. C. Titschmarsh when Hardy went on leave to Princeton.

Cartwright first met Littlewood in June of 1930 when he went to Oxford to examine her for her Doctor of Philosophy degree. The following October she went to Girton College, Cambridge on a three-year research fellowship. She attended a

series of courses by Littlewood on the theory of functions and his seminar that took place in his rooms in Trinity at 5:00 P.M. preceded by a nice tea.

In 1931 Hardy assumed the Sadlerian Chair of Mathematics at Cambridge. Cartwright asked Hardy if he would be offering a seminar similar to the Friday evening sessions she had enjoyed at Oxford. He replied that he would probably come to some arrangement with Littlewood. Soon after, the Lecture List announced a Hardy-Littlewood class. Cartwright recalled[3] that while Littlewood was speaking at the first Hardy-Littlewood class, Hardy came in late, helped himself liberally to tea and began to ask questions. It seemed as if he were trying to pin Littlewood on details, whereas Littlewood was trying to illustrate the main point while taking the details for granted. Littlewood told Hardy that he was not prepared to be heckled and Cartwright does not recall them ever being present together at any subsequent class. Thenceforth, Hardy and Littlewood alternated classes. Littlewood usually did the speaking on his turns and Hardy often invited others to speak during the classes. Eventually, Littlewood ceased to participate, though the class continued to be held in his rooms. The class became known as "the Hardy-Littlewood Conversation Class at which Littlewood was *never* present." Littlewood intermittently held his own unlisted lecture class, which Hardy recommended to several of his students.

During her first years at Cambridge, Cartwright occasionally wrote or spoke with Littlewood on topics pertaining to his courses. She recalls his manner as somewhat unconventional; he was always more ready to talk out of doors. She was occasionally able to catch him on the telephone after her late tea and before his early dinner. He never discussed problems with her at a blackboard, but he would draw imaginary figures on a wall as they walked and talked. Once they began to collaborate, nearly all of their collaboration was done by letter with occasional short discussions of particular points. He never came to her rooms and she does not remember going to his. Most of the letters Littlewood sent Cartwright would have something on the back, often galley proofs.

Cartwright suspects that Littlewood may have had some unspoken rules for their collaboration comparable to those he had with Hardy. Harold Bohr noted that the following four "axioms of collaboration" formed a basis for the Hardy-Littlewood relationship.[4]

1. When one wrote to the other, it was completely indifferent whether what they wrote was right or wrong.
2. When one received a letter from the other, he was under no obligation whatsoever to read it, let alone answer it.
3. Although it did not really matter if they both simultaneously thought about the same detail, still, it was preferable that they should not do so.
4. It was quite indifferent if one of them had not contributed the least bit to the contents of a paper under their common name.

Bohr was of the opinion that, "seldom, or never, was such an important and harmonious collaboration founded on such negative axioms."

Cartwright gives Hardy credit for the formulation of these rules. She contends that such a precise formulation was typical of him. When she questioned Littlewood about these axioms, he replied that the agreement between Hardy and himself was unwritten. He never personally informed Cartwright about the rules, but she surmises that his policy regarding their collaboration was similar. Cartwright was once told that when preparing some material to send to her, Littlewood

"withdrew part of what he had written as being contrary to the rules of collaboration."[5] She suspects that Littlewood may have broken the first rule by criticizing some of her mistakes.

According to Cartwright,[6] one memorable episode between herself and Littlewood concerned an interesting problem that he had presented in his theory of functions class. In particular, Does the function $f(z) = \sum_{n=1}^{\infty} a_n z^n$ analytic for $|z| < 1$, which takes no value more than p times in $|z| < 1$, satisfy $|f(re^{i\theta})| < A(p)(1 - r)^{-2p}$ for $0 < r < 1$, where $A(p)$ is a constant depending only on p and $a_1, a_2, \dots, a_p$? Littlewood gave Cartwright the impression that this was an exciting problem and he would be interested in any significant progress she could make toward its solution.

Meanwhile, in one of Edward Collingwood's classes, Cartwright had learned of Ahlfors' Distortion Theorem. In her bath one night she thought she saw how to apply it to Littlewood's problem. When she later settled down to examine the problem more carefully, she fell into the "usual trap." The proof for the case $p = 1$ was known and she tried to prove the case for $p > 1$ using a modification of that proof rather than her original idea. She sent this attempt to Littlewood. In reply she received a note with the snake in Figure 1. In his note Littlewood explained the common error she had made in her proof.

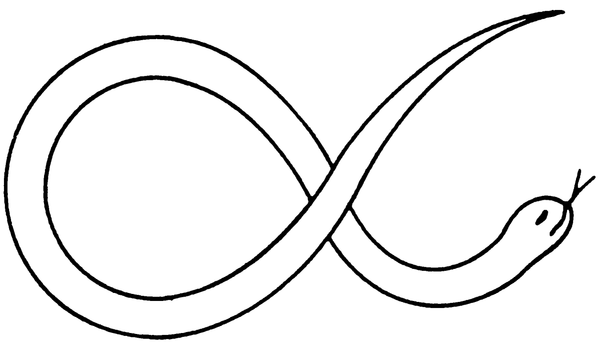
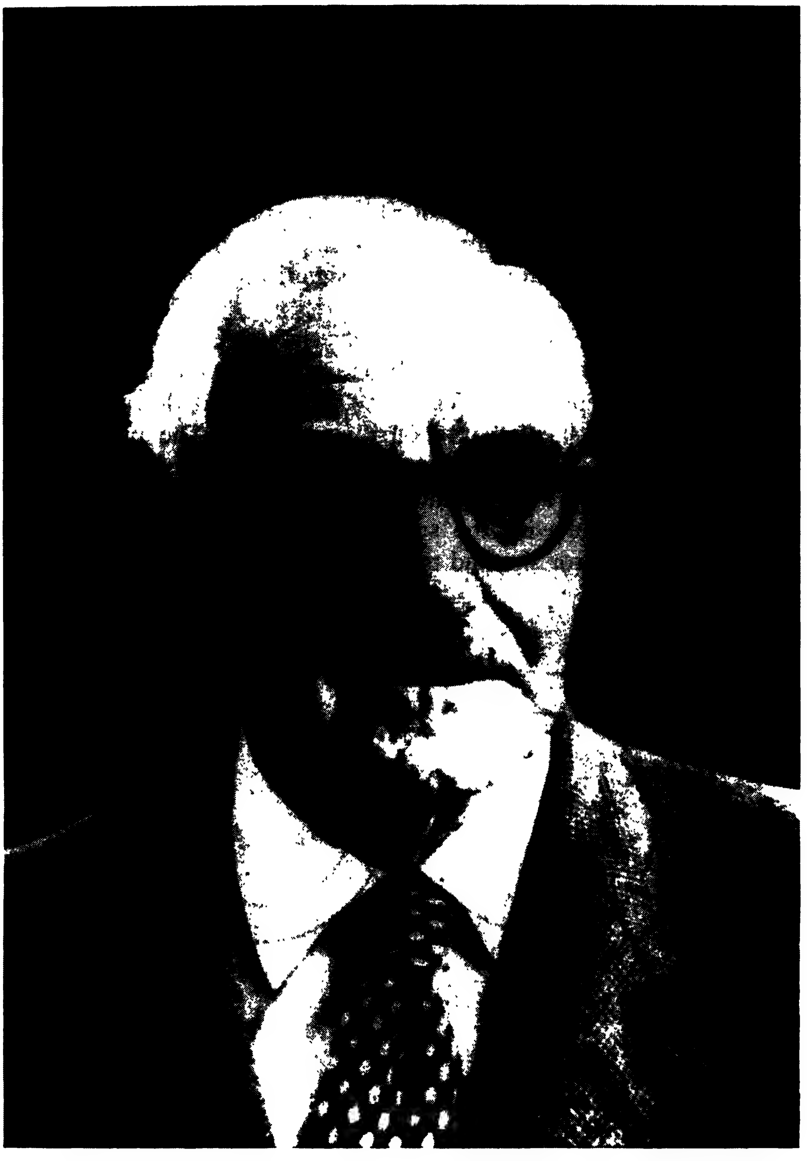


Figure 1

Returning to her original idea, Cartwright sent a new proof to Littlewood using Ahlfors' work. Some time later, while punting with the Vice-Mistress of Girton on the River Cam near Trinity College, Cartwright spotted Littlewood on a nearby bank. As she had received no reply regarding her revised proof, she asked him if he had read her manuscript. He replied, "Have I got to read all that?" She convinced him that he should, and he was impressed with her work.

The polished version of her proof was published[7] and Littlewood referred to the paper in his *Lectures on the Theory of Functions*. Cartwright believes that her work on this problem prompted Littlewood and Hardy to write a joint letter to Girton resulting in a Faculty Assistant Lectureship and the renewal of her research fellowship for a fourth year.

In January of 1938, the Radio Research Board of the Department of Scientific and Industrial Research issued a memorandum requesting the "really expert guidance" of pure mathematicians with "certain types of non-linear differential equations involved in the technique of radio engineering."[8] A copy was sent to the London Mathematical Society. Radio engineers wanted an analysis of the solutions



J. E. Littlewood

to some very objectionable looking differential equations occurring in connection with radar.

The problems that concerned the radio engineers arose from the use of vacuum tubes (thermionic valves) used to control the flow of electricity in the circuitry of transmitters and receivers. Transistors have replaced most vacuum tubes in transmitters; however, vacuum tubes are still sometimes used when very high voltages are present. When encased in ceramic they are very radiation hard and are used in missiles. The type of problem arising from vacuum tubes had great influence on the mathematics of control theory and airplane construction.

The existing techniques available for the explicit analytical solutions of linear differential equations with constant coefficients were the mainstays of early radio engineers. Linear differential equations were used to approximate physical systems and behavior of systems was inferred from the solutions. According to the memorandum, the need had "arisen for a more complete understanding of the actual behavior of certain assemblages of electrical apparatus, including thermionic valves." Nonlinear differential equations were required to obtain a closer approximation of the actual physical system.

Radio engineers hoped mathematicians could provide them with a theory for nonlinear differential equations that had the same relative simplicity as that for linear differential equations. At the time, this request could not be considered completely naïve. Although interest in celestial mechanics inspired considerable research on nonlinear conservative oscillators before 1920, analogous techniques were unavailable for the dissipative systems arising in radio research. Physicists and engineers encountering these systems in their experimental work had done most of the available research. Except for the work of Liapounov and Poincaré, there had been little systematic study of nonlinear differential equations by mathematicians.

Although engineers could compute numerical solutions to their systems, these solutions were of little value when one wanted to know how solutions varied with the parameters. During the early twentieth century numerical solutions were less easy to obtain. Moreover, a large number of numerical solutions are necessary to determine how solutions vary with the parameters of the physical system. The engineers hoped that an analytical approach might provide a more efficient way to analyze their systems.

Unfortunately, mathematical analyses of even the simplest equations representing the physical systems were quite complicated and had few practical applications. The memo acknowledged that it was worthwhile to determine if this were the case in order to prevent the "waste of time and energy spent in pursuit of a will-o'-the-wisp." In the end, complications and variations in the valves themselves convinced radio engineers that experimental methods were more effective than mathematical analysis.

Cartwright says, "Although I myself have helped to develop the general theory and settle certain theoretical problems, I do not think that I have ever produced a result useful for any specific practical problem when it was needed."[9] Though the engineers were unable to apply many of the theories developed by mathematicians such as Cartwright and Littlewood, new problems were arising for automatic control mechanisms involving systems of equations sufficiently similar to those of the radio work for some theory still to be applicable. Resulting mathematical techniques allowed for a headway in the analysis of dissipative systems where progress had previously been rather slow. Such analysis provided part of the foundation to the development of modern dynamical theory.

One of the basic building blocks of early twentieth century circuitry was the triode oscillator illustrated in Figure 2. The objective of the Radio Research Board was to determine which parameter values of the circuit would lead to periodic or almost periodic solutions. It was also essential to determine how the frequency of these oscillations varied with the parameters of the circuit.

The problems appeared interesting to Cartwright. Using the references provided, she began her investigation working back to the research by Appleton and van der Pol in the 1920s. An important paper[10] by van der Pol contained eighty-seven references and was an excellent starting point for her research. The

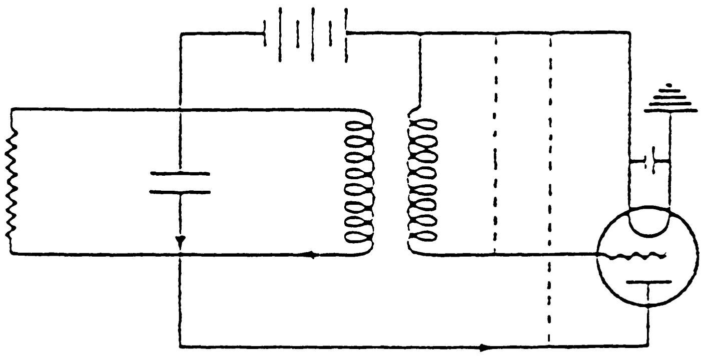


Figure 2

emphasis of van der Pol's article was on phenomena that linear equations could not explain. He referred to several biological and physical research papers such as Volterra's work on interacting species, E. W. Brown's work on the pendulum, and Poincaré's work on celestial mechanics. Cartwright was surprised that van der Pol did not refer to Birkoff's work, Bendixon's work, or Poincaré's paper "*Sur les courbes definies par les equations differentielles*."[11] She says that the well-known ideas of Poincaré on curves and Bendixon on limit cycles have since become fundamental concepts when dealing with several types of nonlinear equations. In addition, she and Littlewood used techniques inspired by Birkhoff's applications of transformation theory to dynamical systems. Her opinion of van der Pol was that, like many authors, herself included, he did not read all the works to which he referred.

It is interesting that Cartwright would concern herself with these applied problems since dynamics was a subject that had not appealed to her at Oxford. Nevertheless, Cartwright intuitively recognized the topological undertones of the problems. She says, "The mathematical problems which have interested me most are apt to turn into topological problems or problems of topological dynamics."[12]

Although she initially knew very little of the dynamical side, Cartwright found the work of the radio engineers to be much more interesting and suggestive than that of mechanical engineers. Radio engineers wanted their systems to oscillate in an orderly way. Similar equations were studied in astronomy and celestial mechanics, but the cycles there are measured in days and years. The speed of radio oscillations, whose cycles are measured in fractions of a second, enables experiments to suggest the behavior of stable solutions quickly and efficiently.

While exploring the literature, Cartwright found several intriguing problems, which she brought to Littlewood's attention. At the time, Cartwright was in the habit of showing Littlewood anything that she thought would interest him. Littlewood dismissed several simple problems, but he found some of van der Pol's conjectures worth investigating.

Cartwright presented Littlewood with the material she had collected because she thought that he might help her understand the dynamics of the situation. He solved or suggested methods of solutions for most of the problems straightaway, but he found certain problems more interesting. Littlewood often thought in dynamical terms, perhaps because of his earlier work in ballistics during the First World War. She says that he often referred to solutions as "trajectories" as if they were the paths of missiles fired from antiaircraft guns.[13] He certainly seemed to

understand the dynamical aspects of the radio work. Van der Pol, recalling a discussion with Littlewood, told Cartwright in "tones of delighted surprise" that all of Littlewood's methods corresponded to physical concepts.[14] Cartwright and Littlewood corresponded with Colebrook, Appleton, and van der Pol in order to gain a better understanding of the situation.

The first problem attacked by Cartwright and Littlewood was that of the amplitude of the stable periodic solution of van der Pol's equation without forcing term:

$$\ddot{x} - k(1 - x^2)\dot{x} + x = 0. \tag{1}$$

Van der Pol had obtained graphical solutions for the cases $k = .1$, 1, and 10 (see Figure 3). Van der Pol's graphs appear to show convergence to a periodic solution with amplitude 2 in all three cases. Littlewood showed that the amplitude was not 2 if k was small and positive. For k small and positive equation (1) has one periodic solution of amplitude $2 + O(k)$. Cartwright and Littlewood succeeded in showing[15] that when k is large all solutions of equation (1), except the trivial solution $x = 0$, converge to a periodic solution whose amplitude tends to 2 as $k \to \infty$. The method they used was elementary, but difficult.

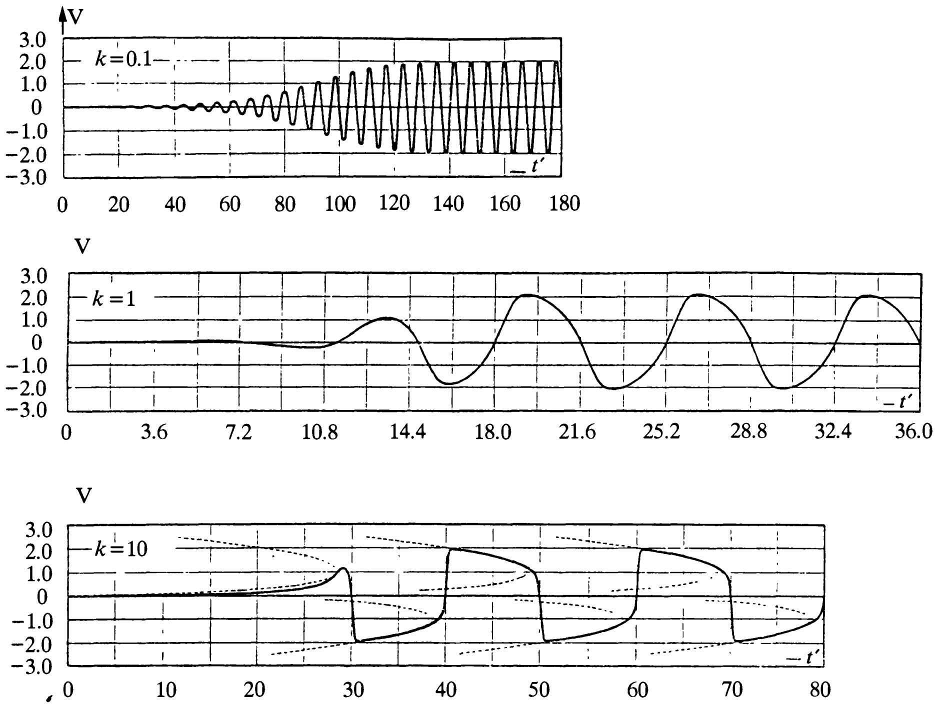


Figure 3

Cartwright and Littlewood next attacked the problem of whether van der Pol's equation with a forced oscillation,

$$\ddot{x} - k(1 - x^2)\dot{x} + x = bk\lambda \cos \lambda t, \tag{2}$$

could have two stable periodic solutions with different periods when k was large. Van der Pol had suggested the phenomenon after he and van der Mark found

experimentally that increasing and decreasing the relative frequency λ of the driving force produced two subharmonics of different orders[16] for the same values of the parameters.

Cartwright and Littlewood focussed a great deal of their attention on van der Pol's equation. They found the equation interesting and suggestive since it seemed to be the simplest type of equation likely to have two stable periodic oscillations with periods prime to one another. They agreed that tackling a really difficult problem first in its simplest form is best.

Most of the early research of Cartwright and Littlewood concentrated on equations with two or more stable periodic solutions. Both considered this problem to be the most analytically interesting problem investigated during their collaboration.[17] Their topological interpretation of this problem led to their paper on fixed point theorems.[18] They left what Littlewood considered the "duller stuff" until later. However, when they explored this material they found that it was more difficult than they had anticipated. In Littlewood's words to Cartwright, "All details have a nasty way of ramifying into difficulties."[19] Their results were published in a paper dealing with second order equations with positive damping, which was much corrected by Littlewood.[20]

Although Cartwright and Littlewood produced some interesting and innovative mathematics, their collaboration was not always harmonious.[21] Cartwright became irritated with Littlewood's incessant changes, even though her own technique was to "polish, polish, polish" before publishing. Cartwright's sense of history was very good and her historical perspective adds an excellent dimension to their papers. When they separated, they agreed that Littlewood would finish the hardest problem and Cartwright would complete the other bits & pieces. Her results appear in a series of papers published in the late 1940s and early 1950s.[22]

A significant portion of the joint work of Cartwright and Littlewood was published under her name alone. This was the case even if the paper was really based on a Littlewood manuscript to which Cartwright had added a bit or filled in the gaps. Littlewood would not let her put his name to any paper not actually written by him. She had to say it was based on joint work with him. There was one exception, the paper on fixed point theorems. Most of the problems they attacked came from the literature of engineering. Their paper on fixed point theorems was an exception as it developed from their treatment of van der Pol's equation.

From 1931 Littlewood suffered from, and was treated for, recurring bouts of mental depression. Cartwright says that his mind was so full of ideas that he found it difficult to rest. She recalls that he seemed exhausted whenever he had completed a collaborative paper and he occasionally complained about their joint papers. Cartwright appreciated the special effort made by Littlewood to have their first paper published in order to support her candidacy for the Royal Society. This joint paper was published in 1945 and was intended as a preliminary survey of their results regarding van der Pol's equation with forcing term and k large.[23] The full proofs of the results presented in the paper were not published until twelve years later. By that time Littlewood had been cured of his depression. Cartwright recalls that in 1958, although he was reluctant, she persuaded Littlewood to go to London to accept the prestigious Copley Medal that the Royal Society had awarded him. She treasures a note from him beginning, "All right, you win."[24] Littlewood probably could not have been convinced to make such a social appearance before his cure.

Cartwright insisted that the two papers published in 1957, those that gave the full proofs of the results announced in their paper of 1945, be published under his

name alone. She may have felt that her contributions had not been significant. However, as Littlewood was leaving for a trip to America, he left some final details of the proof corrections to her. In each manuscript, he affirms that the work was based on joint efforts.

Cartwright and Littlewood's analysis of the van der Pol equation showed that the existence of two stable sets of subharmonics of different orders led to complex topological structures. At the boundary between the domains of attraction of the two sets there exists an intricate fine structure of highly nonstable trajectories. They concluded that the boundary was most likely an indecomposable continuum that, according to Littlewood, is the "dirtiest" thing one can say about a set of points. Except for its instability, this structure closely resembles that of strange attractors. The homeomorphism of the plane defined by following the trajectories through one period maps the set onto itself. Cartwright and Littlewood found this phenomenon to be of great interest since it indicated that the nonstable periodic solutions and subharmonics of the equation corresponded to fixed points of the set under the homeomorphism or an iterate of it. They surmised that existence of these fixed points could be deduced from a suitable fixed point theorem. This led them to investigate fixed point theorems for continua invariant under a diffeomorphism of the plane.

The fixed point theorems needed by Cartwright and Littlewood could not be obtained by the methods of existing algebraic fixed point theory because of the complexity of the continua involved. Fixed point theorems for plane continua that are acyclic but otherwise unrestricted were required. However, they were able to use of some of Birkhoff's results for analytic transformations of the plane into itself with complicated invariant curves. Even reasonable regularity arguments would not allow them to rule out singular invariant curves. Birkhoff had shown the existence of an analytic transformation of the plane into itself with complicated invariant curve K. Cartwright and Littlewood used such a transformation show existence of stable periodic subharmonic motions, unstable motions, and recurrent motions each associated with some subset of K. They were among the earliest mathematicians to associate such transformations with a simple differential equation.

At the time of Cartwright and Littlewood's investigations, Lefschetz, Levinson, Minorsky, and others in the United States also began to explore some of the problems raised by the study of electrical circuits, in particular the boundedness of solutions to the nonlinear differential equations associated with the circuits. Like the English pair, government requests also prompted the Americans. However, whereas the focus of Cartwright and Littlewood was on certain special problems, the Americans were striving for a general and more easily handled mathematical theory. The American emphasis was on amplitude and period. The work of Cartwright and Littlewood emphasized frequencies and parameters. Cartwright speculates[25] that perhaps this difference occurred because engineers became more interested in nonlinearities that were not necessarily either small or large. Meanwhile, Kryloff, Bogoliuboff, and Mitropolsky of the Kiev school were employing averaging methods to analyze equations such as van der Pol's. Until the end of the 19th century, techniques of averaging were used mainly in celestial mechanics. Van der Pol promoted the application of averaging techniques to equations arising in electronic circuit theory. However, rigorous proofs of validity of the method were not available until Fatou provided the first proof of asymptotic validity in 1928. Although they used some foreign research results, Cartwright and Littlewood might have made better use of the work of both the Americans and the Russians if not for the isolation due to the war.

In America, Levinson had obtained results for a piecewise linear equation that were similar to the 1945 results of Cartwright and Littlewood. Cartwright says that a monograph[26] of Levinson's, which combined the work of Kryloff, Bogolieuboff, Denjoy, and Birkoff, considerably influenced the work of herself and Littlewood.[27] Levinson's ideas provided the foundations of a general topological approach to non-autonomous periodic second order differential equations and stimulated the interest of Cartwright and Littlewood in the mappings suggested by their research on the forced van der Pol's equation.

In 1947 Cartwright wrote to Lefschetz at Princeton trying to get a copy of Minorsky's report on nonlinear vibrations. Minorsky was preparing the report for the Office of Naval Research and Lefschetz was listed as head of the project. She included some results of herself and Littlewood in her correspondence. She received the first part of the report and was surprised that it had been classified as "Restricted" and not to be divulged to unauthorized persons. She also received a letter from Lefschetz expressing great interest in her research with Littlewood.

Lefschetz and Mina Rees invited Cartwright to the United States from January to June of 1949 to lecture on nonlinear differential equations. She spent three weeks at Stanford with Minorsky, one week at UCLA with John Curtis, and the rest at Princeton. Cartwright gave a series of lectures[28] at Princeton that provides a detailed overview of her joint work with Littlewood. While at Princeton, she was officially a consultant under the Office of Naval Research, not a professor. At the time, Princeton did not have women professors. Cartwright got on well with all those she met, but Bochner was not on speaking terms with Lefschetz so she never saw him. She also learned that American academicians tended to be a bit casual. John Tukey told her to feel free to put her feet on the table in the seminar room. Wearing a skirt, not trousers, she refrained. Cartwright also learned that if Lefschetz stopped asking questions of a visiting lecturer for five minutes, he was asleep.

Cartwright suspects that Littlewood might not have read much of the literature from which their problems were extracted. She recalled that Littlewood had once advised her not to pay too much attention to the existing literature on a problem. He surmised, "If previous writers had failed to solve it, it was probably because they had tried a wrong method."[29] Consequently, Cartwright and Littlewood approached their problems using methods that were quite different from those of the radio engineers. Their innovative approach led Cartwright and Littlewood to the conclusion that using both topological and analytical methods was indispensable in the study of differential equations. According to Cartwright, "We should widen our conception of topological methods to include all those now claimed as such by topologists and combine them with analytical methods to obtain results more quickly; we need a clarification of existing topological results in terms of differential equations... ."[30] Cartwright also stressed that the topological interpretation of a problem can give one valuable insight into the qualitative behavior of solutions, even when one is unfamiliar with or does not plan to use topological methods in their analysis.[31]

Cartwright found it remarkable that the experimental results of van der Pol and van der Mark guided her and Littlewood to so many unsolved problems in topology.[32] Several papers[33] that came from the collaboration of Cartwright and Littlewood are among the earliest fully rigorous works in large parameter theory, i.e., relaxation oscillations. Work such as theirs clearly provided an impetus to the development of the modern theory of dynamical systems. The subject has expanded significantly since their work was completed. Modern techniques have

Dame Mary Cartwright
(Courtesy of the Mistress and fellows of Gurton College)

enabled contemporary mathematicians to carry through a more thorough analysis of the asymptotic behavior exhibited by solutions to the forced van der Pol equation.[34] Nonlinear oscillation theory has led to the development of radio, radar, and laser technology, and investigations into the van der Pol equation have played a vital role in both the theory of relaxation oscillations and bifurcation theory.

REFERENCES

1. S. Smale, *Differential and Combinatorial Topology*, pp. 63–80, Princeton, 1965.
2. M. L. Cartwright, *Religion and the Scientist*, Mervyn Stockwood, ed., p. 23, SCM Press, London, 1959.
3. M. L. Cartwright, John Edensor Littlewood, FRS, FRAS, Hon FIMA, *Bull. Inst. Math. Appl.* **14** (1978), 89.
4. H. Bohr, *Coll. Math. Works* **1**, p. xxviii. Dansk Mathematisk Forening, Kobenhavn, 1952.
5. M. L. Cartwright, John Edensor Littlewood . . . , p. 90.
6. M. L. Cartwright, Some exciting mathematical episodes involving J.E.L., *Bull. Inst. Math. Appl.* **12** (1976), 201.
7. M. L. Cartwright, Some inequalities in the theory of functions, *Math. Annalen* **111** (1935), 98–118.
8. A note on certain types of non-linear differential equations involved in radio engineering, *Math. Gazette* **22** (1938), 217–218.
9. M. L. Cartwright, Mathematics and thinking mathematically, *Amer. Math. Monthly* **77** (1970) 23.
10. B. van der Pol, The nonlinear theory of electric oscillations, *Proc. Inst. Radio Eng.* **22** (1934), 1051–1086.
11. M. L. Cartwright, Some points in the history of the theory of nonlinear oscillations. *Bull. Inst. for Math. and its Appl.* **10** (1974), 329.
12. M. L. Cartwright, From non-linear oscillations to topological dynamics, *Jour. London Math. Soc.* **39** (1964), 1931.
13. M. L. Cartwright, Mathematics and thinking mathematically, 22.
14. M. L. Cartwright, John Edensor Littlewood . . . , 88.
15. M. L. Cartwright, Van der Pol's equation for relaxation oscillations, *Contributions to the Theory of Nonlinear Oscillations II; Annals of Mathematical Studies* **2**, pp. 3–18, Princeton, 1952.
16. If a differential equation with a forcing function of period P has a periodic solution with period nP, where n is a positive integer, then that oscillation is called a subharmonic of order n.
17. M. L. Cartwright, Some points in the history . . . , 333.
18. J. E. Littlewood and M. L. Cartwright, Some fixed point theorems, *Annals of Math.* **54** (1951), 1–37.
19. M. L. Cartwright, From non-linear oscillations . . . , 196.
20. J. E. Littlewood and M. L. Cartwright, On nonlinear differential equations of the second order II. The equation

$$\ddot{y} + kf(y)\dot{y} + g(y,k) = p(t) = p_1(t) + kp_2(t);\ k > 0,\ f(y) \geq 1,$$

Annals of Math. **48** (1947), 472–494; Addendum **50** (1949), 504–505.
21. B. Bollobás, ed. *Littlewood's Miscellany*, p. 13. Cambridge, 1986.
22. The papers by M. L. Cartwright based on joint work with J. E. Littlewood are, Forced oscillations in nearly sinusoidal systems, *Jour. Inst. Elec. Engrs.* III **95** (1948), 88–96; On nonlinear differential equations of the second order III. The equation

$$\ddot{x} - k(1 - x^2)\dot{x} + x = pk\lambda\cos(\lambda t + \alpha),$$

k small and near 1, *Proc. Cambridge Phil. Soc.* **45** (1949), 495–501; Forced oscillations in nonlinear systems, *Contributions to the Theory of Nonlinear Oscillations. Annals of Math. Studies* **20**, pp. 149–241, Princeton, 1950; Forced oscillations in nonlinear systems, *Jour. Research Nat. Bur. Standards* **45** (1950), 514–518; and Van der Pol's equation for relaxation oscillations, *Contributions to the Theory of Nonlinear Oscillations II. Annals of Math. Studies* **2**, pp. 3–18, Princeton, 1952.
23. M. L. Cartwright, Some inequalities . . . , *Math. Ann.* **111** (1935), 98–118.
24. M. L. Cartwright, John Edensor Littlewood . . . , 89.
25. M. L. Cartwright, Some points . . . , 332–333.
26. N. Levinson, Transformation theory of nonlinear differential equations of the second order, *Annals of Math.* **45** (194), 723–737.
27. M. L. Cartwright, Some points in the history . . . , 333.
28. M. L. Cartwright, Forced oscillations in nonlinear systems, *Contributions to the Theory of Nonlinear Oscillations. Annals of Math. Studies* **20**, pp. 149–241, Princeton, 1950.
29. M. L. Cartwright, John Edensor Littlewood, . . . , 88.
30. M. L. Cartwright, Topological problems of nonlinear mechanics, *Konferenz über Nichtlineare Schwingungen, 1964, Berlin, Abh. Deutsch. Akad. Wiss. Berlin. Kl. Mathematik Physik, und Technik*, #1, p. 141, 1965.
31. M. L. Cartwright, E. T. Copson, and J. Grieg, Non-linear vibrations, *Advancement Sci.* **6** (1949), No. 21, 64–75.

32. M. L. Cartwright, Non-linear vibrations: A chapter in mathematical history, *Math. Gazette* **36** (1952), 81.
33. J. E. Littlewood, On nonlinear differential equations..., *Acta Math.* **97** (1957), 267–308; J. E. Littlewood, On nonlinear differential equations..., *Acta Math.* **98** (1957), 1–110; J. E. Littlewood and M. L. Cartwright, On nonlinear differential equations..., *Annals of Math.* **48** (1947), 472–494; Addendum **50** (1950), 504–505; M. L. Cartwright, Van der Pol's equation..., *Annals of Math. Studies* **2**, pp. 3–18, Princeton, 1952.
34. See, for example, M. Levi, Qualitative analysis of the periodically forced relaxation oscillations, *Mem. Amer. Math. Soc.* **32** (1981), 244; J. Guckenheimer, Dynamics of the van der Pol equation, *IEEE Trans. Circuits and Systems* **27** (1980), 983–989; J. Guckenheimer, Symbolic dynamics and relaxation oscillations, *Physica D. Nonlinear Phenomena* **1** (1980), 227–235.

Department of Mathematics
Providence College
Providence, RI 02918

Where math teachers meet and eat in Maastricht, Netherlands.
Photo by Sarah MacMillen.

Contributed by Evan Bonner,
Susquehannah Valley High School, Conklin, NY.

What Are Infinitesimals and Why They Cannot Be Seen

Roman Kossak

Visualizing mathematical structures is a broad and important subject in mathematics education. Consider, for example, the difficulty of seeing a number system that includes infinitesimals. This might seem an obvious observation since infinitesimals are infinitely small, but that is not the point I wish to make here. In mathematics we have tools that allow us to think visually about objects far beyond direct physical perception. We can see the infinite. We use our natural intuition of the geometry of three dimensional space as a starting point for constructions of abstract mathematical structures, such as higher dimensional vector spaces, non-euclidean geometries, and topological spaces of various kinds. Those structures often do not seem to have direct interpretations in the physical universe we know, but nevertheless we can picture them in the mind's eye. With simple images we can illustrate, quite accurately, more abstract concepts.

Precision and formal rigor are essential in mathematics, but these qualities apply only to the final products of mathematical activity. When we "do" mathematics, we rarely think formally. We do not limit ourselves to computations and rigorous proofs. Rather, mathematicians work with pictures—visual representations of mathematical structures. The same pictures also seem indispensable in learning mathematics. Some mathematical structures, however, cannot be visualized easily. A level of abstraction and set theoretic methods are necessary to prove that certain structures even exist. A theorem of Stanley Tennenbaum's [**5**], published in 1959, states that certain mathematical structures are very complex. The notion of complexity, as used in Tennenbaum's theorem, has precise meaning, and I will say more about it later. At this point let me just mention that the theorem excludes the possibility of a simple presentation of some structures, and the number system of nonstandard analysis is one of them.

Nonstandard analysis was created (or perhaps recreated) by Abraham Robinson [**4**], who reintroduced infinitesimals to mathematics using the formal apparatus of mathematical logic. Since then, much has been written on the theory of infinitesimals and their applications. Part of the effort was directed towards utilizing some of the new methods and techniques in teaching. Nonstandard analysis offers great simplicity in defining basic concepts. Think, for example, of the following definition of the derivative: $y = f(x)$ is differentiable at a point a, and b is the derivative of $f(x)$ at a, if for every number dx infinitesimally close to 0,

$$\frac{dy}{dx} = \frac{f(a + dx) - f(a)}{dx}$$

is infinitesimally close to b. The derivative becomes a quotient of two numbers![1]

[1]Technically speaking, $f'(a)$ is the real number that is infinitesimally close to the quotient dy/dx.

Consequently, many of the informal arguments of differential and integral calculus become precise.

Attractive calculus textbooks based on nonstandard analysis have been written ([1], [3]). In [1] we read "A most natural place for Robinson's insight is a next (and possibly final) point in the evolution of the teaching of calculus. We can now develop calculus using infinitesimals and enjoy all of their simplicity and intuitive power, yet at the same time work in a mathematically precise and rigorous atmosphere." But somehow the new trend has not gained enough popularity to revolutionize the teaching of calculus. Perhaps, like almost everything else, nonstandard analysis has been overshadowed by the reform of calculus education based on the power of Maple and Mathematica. Or perhaps the reason lies also somewhere else. As I said before, the problem with infinitesimals is that we do not have a good model to look at when we think about them.

The number system including infinitesimals can be introduced axiomatically. We can declare that infinitesimals do exist, and we can list all their properties that are needed in order to proceed with the development of calculus. If we teach calculus this way, we have to assume that the notion of an infinitesimal is in some sense natural and intuitive, and if it is not, it becomes natural and intuitive during the course of the study. But if the question, "What are the infinitesimals?" arises, we do not have a really good answer, except for something like, "Mathematical logic takes care of that."

In this article, I want first to remind the reader why it would be good to have infinitesimals at hand when we teach calculus. Then I want to comment briefly on the development of the real number system. I will describe some structural properties of a number system with infinitesimals, and, finally, I will elaborate on Tennenbaum's theorem.

1. CONTINUITY OF ELEMENTARY FUNCTIONS. When we learn about graphs of polynomial functions, we are asked first to plot some points to sketch the graph of the function $y = x^2$. When the shape of the graph emerges, we join the points with a continuous line. But how do we know that the line will be continuous? The claim can be justified by plotting a larger number of points, or even better, with a computer or a graphic calculator, producing a curve that looks continuous. This, however, is not a good example of precise mathematical reasoning. After all, what we are doing is only an approximation consisting of a set of points that might be large but still is finite (I am afraid many students do not know that). So, the inquisitive student could ask, "What if we were not careful enough?" There might be points on the graph that break the regular pattern. For example, what about squares of irrational numbers? How do we compute those? The answer is not easily available. What we see on the computer screen is very convincing, but when dealing with functions as elementary as $y = x^2$, we should be able to provide good explanations without employing sophisticated technology. Clearly, the continuity of $y = x^2$, and other polynomial functions, should follow directly from elementary properties of addition and multiplication.

But what is continuity? Well, we can say that a function $y = f(x)$ is continuous at a point a, if $f(x)$ is defined in some open interval containing a, and for every $\varepsilon > 0$ there exits $\delta > 0$ such that for all x in the domain of $f(x)$, if $|a - x| < \delta$ then $|f(a) - f(x)| < \varepsilon$. This is a perfectly good definition, and a simple proof of continuity of $y = x^2$ can be easily supplied. However, such a proof seems completely useless in the didactic process. Students are puzzled by the logical complexity of the argument. And indeed, the argument is unnatural. This might be the

reason why a discussion of continuity of polynomial functions is almost completely eliminated from brief calculus texts. What is offered instead, is the illustration of the concept of continuity using piecewise definable functions. I have found, much to my surprise, that it is this part of the introduction to calculus that students find particularly difficult. I think now that this is quite justifiable. Although a well-trained student has absolutely no problems with this material, beginners have a hard time trying to figure out what continuity is. On the one hand, the intricacy of the mathematical apparatus (limits) suggests that we are dealing with a profound and complex subject. On the other hand, the examples illustrating the concept (piecewise defined functions) are very simple. Students at this point are supposed to know how to sketch graphs of elementary functions, so they can answer all the questions just by inspecting the graphs. That creates confusion.

The continuity of the quadratic, and other elementary functions, is an important topic that should not be glossed over. Can something be done about it? Take a look at the following proof.

We will say that a function $y = f(x)$ is continuous at a point a, if $f(x)$ is defined in some open interval containing a, and for every point b that is infinitesimally close to a, $f(b)$ is infinitesimally close to $f(a)$. Let us prove now that the quadratic function is continuous. Suppose b is infinitesimally close to a. Then $b = a + dx$, where dx is infinitesimally close to 0. Hence

$$b^2 = (a + dx)^2 = a^2 + 2a(dx) + (dx)^2.$$

Now, since $2a(dx) + (dx)^2$ is infinitesimally close to 0, b^2 is infinitesimally close to a^2, and the proof of continuity of $y = x^2$ is completed. Our task now will be to explain what makes this argument correct. Before we do this, let me note that, assuming one knows what it means for two numbers to be infinitesimally close, the preceding proof is much easier than anything we can do using the $\varepsilon - \delta$ approach. It uses elementary properties of arithmetic operations, and we could risk the claim that it really explains why $y = x^2$ is continuous.

2. SEEING THE NUMBER SYSTEM. A straight line with an infinite collection of evenly distributed points is the illustration that probably all of us see when we think of integers. Seeing a line with points on it, however, is not enough; we need to know how numbers are added and multiplied. Simple illustrations can be used to explain the rules of the arithmetic of positive integers. The arithmetic of negative numbers requires more work, but it can be done convincingly. Now comes the time to introduce the rationals. We represent fractions as points on the same line that serves as an illustration for the integers. Again, it takes a bit of work to explain how to add and multiply fractions and how arithmetic operations relate to the geometric structure of the line. Then we introduce the coordinate system, and it seems as if we could begin a presentation of calculus.

Between any two rational numbers there is another rational number. Any segment of the number line can be partitioned into arbitrarily small subintervals. Rational numbers form a *dense* structure. So, in fact, to see the picture, I have to have to draw a *continuous* line. But now comes a surprise. Let O be the circle, centered at the origin of the number line, whose radius is equal to the length of the diagonal of the unit square. We can show that O and the rational number line have no common points. The rational number line has gaps in it! That is unexpected and rather difficult to see. The line seemed continuous only a moment ago. In the classroom, we should take a moment and explain carefully what is

happening here. We rarely do that. If this important moment is overlooked, then there is almost no point in discussing later more theoretical aspects of calculus such as, for example, the intermediate value theorem for continuous functions.

It takes an effort to build an image of the real number line for one's private use. We have to convince ourselves that although there are (many!) irrational numbers, and in fact the rational line looks like a sieve, the intuition we developed while thinking about rational numbers is still helpful in understanding the reals.

We have to spend some time thinking about the arithmetic of the real numbers. How to compute π^2? What is $\sqrt{2}^{\sqrt{2}}$? Is it rational? The answers are not immediate, and this difficulty should not be concealed. We do not compute with real numbers; we use their rational approximations, and that somehow suffices for all practical applications. Hence, we need to address the question, "Why do we bother with irrational numbers at all?" There are many answers. I usually say, "We can compute anything we want without irrationals, but we could never understand geometry and calculus without them."

The reals form a vastly more complex structure than the rationals, but still, to follow an elementary calculus course and to understand the applications, the vague picture consisting of a rational number line supplemented with some known irrationals stuffed somewhere there is quite sufficient. The status of infinitesimals is quite different.

3. THE ARITHMETIC OF INFINITESIMALS.

Although the rationals form a dense set, still we managed to squeeze the irrationals in. Now it is time to try to squeeze in even more.

Imagine that 0 is surrounded by a set of numbers dx such that $a < dx < b$ for every negative real number a and every positive real number b. We call the set of such numbers dx the *monad* of 0. Thus, 0 is separated from all other real numbers by its monad. The elements of the monad of 0 are called *infinitesimals*. How many infinitesimals are there? How should they be added and multiplied?

First of all, we want to assume that all elementary properties of addition and multiplication of real numbers are still valid for all numbers including the infinitesimals. Thus, for example $dx + dx' = dx' + dx$ for all infinitesimals dx, dx'. We also want the numbers to be linearly ordered in such a way that if dx is an infinitesimal and $dx > 0$, then $(-1)dx < 0$.

It follows that for every number b the set of numbers of the form $b + dx$, where dx is infinitesimal, forms a monad of b, i.e., the set of numbers infinitesimally close to b. It is clear that $b + dx$ is infinitesimally close to b; conversely, if a is infinitesimally close to b, then $a - b = dx$, for some infinitesimal dx, hence $a = b + dx$.

Many other elementary properties of infinitesimals can be shown in a similar fashion. As exercises, prove that if a is infinitesimally close to b, then a and b have the same monads, that the sum and a product of two infinitesimals is infinitesimal, and that, if dx is infinitesimal and r is a real number, then $r \cdot dx$ is infinitesimal. These are all the facts we need to complete the proof of the continuity of $y = x^2$.

Thus, the proof of continuity of $y = x^2$ is finished, but many obvious questions arise. Have we proven anything? Notice that our considerations were hypothetical. All we said was, "If there are infinitesimal numbers, and if they have the properties we need, then the sum of two infinitesimals is infinitesimal, etc." But

what if there are no infinitesimals? What if there are no extensions of the real number system with the required properties? What if, speaking in logical terms, the requirements are inconsistent?

The proof of continuity of the quadratic function requires only very elementary properties of infinitesimals. Similar arguments work for other polynomial functions. But what about more complex elementary functions? What is $1/dx$? How can one compute $\sin(dx)$ and $\cos(dx)$? What does the trigonometry of infinitesimals look like?

As I mentioned before, the answer to all these questions is provided by mathematical logic. There are extensions of the real number system in which we can abandon the $\varepsilon - \delta$ formalism in favor of proofs and constructions involving infinitesimals. The problem is that those extensions are neither unique nor easily defined. In the next section I will try to describe what they look like.

4. ATTEMPTING TO SEE THE INFINITESIMALS. If dx is an infinitesimal, what should $1/dx$ be? Suppose that dx is positive. We know that $dx < b$ for every positive real number b. In particular $dx < 1/n$ for every positive integer n. Since the ordinary rules of multiplication apply, we then must have $n < 1/dx$, for every positive integer n. Similarly, the reciprocal of a negative infinitesimal must be smaller than all negative integers. This means that our extended number system, along with infinitesimals, must also include infinitely large numbers.

Thus, the reciprocal of an infinitesimal is an infinitely large number. Similarly the reciprocal of an infinitely large number has to be infinitesimal. The function $f(x) = 1/x$ provides a one to one correspondence between the positive elements of the monad of 0 and the set of infinitely large numbers (which we could call the monad of infinity). Thus, instead of trying to describe the structure of the monad of 0 directly, we can concentrate on the structure of infinitely large numbers.

The set of infinitely large numbers contains an extension of the set of natural numbers; we call elements of this extension *nonstandard integers*. Nonstandard integers are not only a technical curiosity. They play an important role in the development of nonstandard integral calculus, where infinite sequences and series are replaced by sequences and series of nonstandard integer length, and counting techniques are used instead of limits.

The structure we are interested in is $(\mathbf{R}^*, \mathbf{Z}^*, 0, 1, +, \cdot, <)$, where $\mathbf{R}^*$ is all numbers, including infinitesimals and infinitely large numbers, and $\mathbf{Z}^*$ is the set of the integers of $\mathbf{R}^*$, including the standard and nonstandard integers.

Let us examine what $(\mathbf{N}^*, 0, 1, +, \cdot, <)$, the positive part of the integers of $\mathbf{R}^*$, should look like. We know that $\mathbf{N}^*$ must contain some number c such than $n < c$ for all standard integers n. It is not difficult to imagine a simple structure consisting of all positive integers and one number c on top of it. Think, for example, of the set of numbers of the form $1 - 1/n$, for $n = 1, 2, \ldots,$ with the natural ordering, and identify c with 1. But we have to do more.

Obviously, for every standard integer k, the number $c + k$ must be nonstandard. This gives us infinitely many nonstandard integers. Is that enough? Not yet. For every standard k, we have $c + k < c + c = 2c$; hence we need a nonstandard integer $d = 2c$ that is greater than all numbers of the form $c + k$, where k is standard.

Together with d we get all numbers of the form $d + k$, for standard k, and these numbers are greater than numbers of the form $c + k$, for standard k. Similarly, we need numbers of the form $c - k, d - k$, etc., and one can easily check that for all standard k, l we have $c + k < d - l$. Then come $3c, 4c$, and all

other numbers of the form nc, where n is a positive integer. The structure is already quite rich, but still is "visible."

Now it is time to ask about c^2. Clearly nc, for finite n, must be smaller than $c \cdot c$, hence c^2 is not among the elements we have got so far. We need more, in fact many more, elements. The next step would be to take all elements of the form $p(c)$, where $p(c)$ is a polynomial with (standard) integer coefficients, and whose leading coefficients is positive. The set is easy to define, but we also need to decide what the structure is going to be: we need to define addition and multiplication for nonstandard numbers obtained this way. This turns out to be easy. We define $p_1(c) + p_2(c)$ to be $p_3(c)$, if $p_1(x) + p_2(x) = p_3(x)$ in $\mathbf{Z}[x]$, and similarly for multiplication. Unfortunately, we cannot stop here. Since, for every standard n, $k^n < 2^k$ for all k that are sufficiently large, and c certainly is sufficiently large, we must have $c^n < 2^c$ for every standard n. Since we want to be able to use exponentiation in our system, we have to keep adding more nonstandard integers. But even if we add $e = 2^c$ to our structure, together with all elements of the form $p(e)$, and if we iterate this process infinitely many times, the structure we get is still too small.

Our aim is to obtain a structure for which the following *transfer principle* holds. If A is a formal statement in the language of arithmetic (without nonstandard parameters), and if the elements of $\mathbf{Z}^*$ behave according to A, then A is also a true statement about the standard numbers. In other words, we want $\mathbf{Z}^*$ to have the same arithmetical properties as $(\mathbf{Z}, 0, 1, +, \cdot, <)$.[2]

A structure $\mathbf{Z}^*$ with this property can be built with the help of mathematical logic. $\mathbf{Z}^*$ is very complex. One of the reasons is that every number system for infinitesimal calculus must contain uncountably many nonstandard integers. Thus, if we'd like to visualize the structure of the nonstandard integers, we would have to consider problems of set theoretic nature: "What is the continuum and what is its structure?" Set theory teaches us that one cannot expect simple answers in this matter. But still, there might be some easier way to visualize some countable structures that could serve as informal models of nonstandard integers. The role of those models would be similar to the role of the rationals in the construction of the reals.

Structures that could be used for this purpose were known long before Robinson's discovery. In 1934, Thoralf Skolem described a construction of a structure of the form $(M, +, \cdot, 0, 1)$ that extends the standard integers and contains nonstandard numbers. Such structures are called *nonstandard models of arithmetic*. There are countable nonstandard models of arithmetic for which the transfer principle holds. Such structures can be used as approximations to a nonstandard uncountable model formed by positive integers of a number system for infinitesimal calculus. Tennenbaum's theorem shows that even countable nonstandard models have to be very complex. As I will try to show in the next section, they are certainly too complex to serve as a visual aid.

5. TENNENBAUM'S PHENOMENON.

Let $\mathbf{Z}[x]^+$ denote the set of polynomials with (standard) integer coefficients whose leading coefficient is positive. $\mathbf{Z}[x]^+$ has naturally defined operations of addition and multiplication. We can also define the

[2] A definition of a "formal statement in the language of arithmetic" is necessary here. For the purpose of our discussion it is enough to think of true statements about the standard numbers as theorems of elementary number theory.

relation: $p < q$ if and only if either the degree of p is smaller than the degree of q or the degrees are the same and the leading coefficient of p is smaller than the leading coefficient of q. $\mathbf{Z}[x]^+$ with operations and relation described forms a structure that can serve as an approximation to a nonstandard model. As we have seen in the previous section, if c is nonstandard, then the nonstandard model contains the set $P(c) = \{p(c)\colon p \in \mathbf{Z}[x]^+\}$ and many other elements (the arguments used in the previous section concerned nonstandard integers in a number system of nonstandard analysis, but they apply as well to any nonstandard model of arithmetic).

There are two points that I want to make now. The first is that $\mathbf{Z}[x]^+$ is a simple mathematical structure, and the second is that $\mathbf{Z}[x]^+$ is isomorphic to $P(c)$, hence the latter is a simple structure as well.

The notion of simplicity (and complexity) of a structure is borrowed from the theory of computations. We say that a mathematical structure is *simple* if it is isomorphic to a structure whose operations can be performed by a computer program. It would be difficult to claim that a structure that is simple according to the preceding definition is also easy to visualize. On the other hand, I believe it can be argued convincingly that a structure that is not simple is very difficult to see.

The polynomials in $\mathbf{Z}[x]^+$ can be effectively enumerated. This means that we can make a list $p_0, p_1, \dots, p_i, \dots$ of all polynomials in $\mathbf{Z}[x]^+$, and we can write a computer program that will produce the i-th polynomial on that list given the input i. Also, since addition and multiplication of polynomials are defined effectively, two other programs can be written. The inputs for both programs are pairs of numbers (i, j). The output of the first program on the input (i, j) is the index k, such that $p_k = p_i + p_j$. The output of the second program on the same input is the index l, such that $p_l = p_i \cdot p_j$. The three programs make $\mathbf{Z}[x]^+$, with its arithmetical operations, simple.

Now, why is $P(c)$ isomorphic to $\mathbf{Z}[x]^+$? We have to be a bit careful. If, instead of a nonstandard c, we take a standard number n, then, for example, $p(n) = q(n)$, where $p(x) = x^2$ and $q(x) = nx$, hence two distinct elements of $\mathbf{Z}[x]^+$ correspond to the same element of $P(n)$. This, however, does not happen in $P(c)$. Let p, q be given polynomials. Every equation of the form $p(x) = q(x)$ has a finite number of solutions, hence there is an integer k such that for all $n > k, p(n) \neq q(n)$. According to the transfer principle, for all elements n of the nonstandard model, if $n > k$, then $p(n) \neq q(n)$. Clearly, $c > k$, hence $p(c) \neq q(c)$.

In this manner, we have effectively presented $(P(c), 0, 1, +, \cdot, <)$—a part of the structure of the nonstandard model. Unfortunately, as we have seen, for every nonstandard number c, $P(c)$ is a proper subset of the model.

The model-theoretic construction of a nonstandard model M proceeds as follows. Roughly speaking, M can be represented as a set of elements of the form $f(c)$, where c is a nonstandard number and $f(x)$ ranges over the family of all arithmetically defined functions (including polynomial functions, exponential functions, and more). The problem is to decide for which functions f and g, $f(c)$ and $g(c)$ represent the same element. We know that if $f(n) \neq g(n)$ for all standard integers n greater than some k, then $f(c)$ and $g(c)$ represent distinct nonstandard integers. In other cases, the problem of deciding whether $f(c) = g(c)$ becomes delicate. One of the reasons is that our knowledge about *arbitrary* arithmetic functions is rather limited. Think for example of the function $f(n) =$ the least $k > n$ such that k and $k + 2$ are prime. But the real difficulty here is not insufficient knowledge. The obstacle is of a more fundamental nature.

Nonstandard models of arithmetic, and universes of nonstandard analysis, can be built in many ways. There is a vast variety of nonisomorphic structures of this kind. The necessary tools—ultrapowers—can be found in any textbook on model theory. The context in which this is done is abstract and set theoretic. We can study these structures in great detail, but this requires some training. The methods we use were not available to mathematicians of the XIX-th century, not to mention Leibniz and Newton. Tennenbaum's theorem, which we present now, shows that all of these structures are complex not only because we made them so, but because they have to be.

Suppose M is a countable nonstandard model of arithmetic. Suppose that elements of M can be listed, i.e., $M = \{a_1, a_2, \dots\}$. Suppose that $F(x, y)$ is any computable function, i.e., a function for which there exists a computer program that computers $F(m, n)$ on the input (m, n). Tennenbaum's theorem implies that there are integers m and n such that $F(m, n) = k$ and $a_m + a_n \neq a_k$ (addition here is performed in M). In other words, the theorem says that no computable function can represent the addition of a nonstandard model. Every such function $F(x, y)$ produces wrong answers on some elements of M if we try to use it as a calculator for adding numbers in M. (It also follows that every calculator must in fact produce wrong answers in infinitely many cases.) Tennenbaum's theorem says also that no computable function can represent the multiplication of M.

Thus, in conclusion we now can say: no structure of nonstandard integers is simple.

Tennenbaum's theorem is not difficult to prove, given some preparation in recursion theory and model theory. A detailed proof with a discussion is given, for example, in **[2]**. I believe that the theorem and its philosophical consequences (the problem of existence of mathematical objects) and perhaps even practical consequences (the teaching of calculus) deserve a much wider recognition.

ACKNOWLEDGMENT. I would like to thank Jim Henle and Jim Schmerl for comments on drafts of this article.

REFERENCES

[1] J. Henle, E. Kleinberg, *Infinitesimal calculus*, MIT Press, 1979.
[2] R. Kaye, *Models of Peano Arithmetic*, Oxford Logic Guides 15, Oxford University Press, 1991.
[3] J. Keisler, *Foundations of infinitesimal calculus*, Prindle, Weber and Schmidt, 1976.
[4] A. Robinson, *Non-standard analysis*, North Holland, 1966.
[5] S. Tennenbaum, Nonarchimedean models for arithmetic. *Notices of the American Mathematical Society*, 270, 1959.

Department of Mathematics and Computer Science
Bronx Community College / CUNY
University Avenue and West 181 Street
Bronx, NY 10453
rkobb@cunyvm.cuny.edu

A Case Study in Mathematical Research: The Golay-Rudin-Shapiro Sequence

John Brillhart and Patrick Morton

1. INTRODUCTION. The case study we are presenting here is a re-creation of our original investigation into the Golay[1]-Rudin-Shapiro sequence [3]. We are particularly fond of this investigation, because of its unexpected simplicity and elegance. It contains a nice balance between reasonable and thought-provoking questions that lead us through the development of the subject and answers that arise from examining pertinent data as we go along. These answers lead in turn to more questions, etc. Our main purpose in re-creating this investigation is to show the evolution of questions and ideas that originally led us to our results. Thus, we are especially interested in highlighting some of the stances that mathematicians take in the middle of their work.

The standard and time-honored practice in mathematics—to erase all hint of the development of a subject or proof—usually makes it hard for students to see into the minds of mathematicians at work. Theorems and arguments seem to come from nowhere. Very seldom in textbooks or in research papers is there a hint of the original questions that motivated the researchers, or what special turns their understanding took in the middle of developing their subject. For us, that is one of the most exciting things about doing mathematics. We hope that students will see that the thought processes mathematicians engage in are much the same as the normal human process of asking questions and being alert to hints suggested by the subject itself. This questioning and following leads are at the heart of successful mathematical endeavors.

A secondary purpose of this paper is to provide an introduction to the subtleties of the Golay-Rudin-Shapiro sequence, a sequence that has motivated many interesting developments in the last 25 years. (See [1], [4], [5], [7], and the references contained in those papers.)

We kept an undergraduate audience in mind as we wrote this study. We envision it being used for its examples in elementary real analysis: for the empirical investigation of maxima and minima, arguments involving limit points, lim sups and lim infs, experimenting with inequalities, even for the experience of a frustrated attempt to solve a problem. This paper also serves as an introduction to topics that more advanced students can read in [4]. We think it might be suitable as an introduction to research methods for students involved in summer research programs or independent study.

[1]The sequence was originally named after Shapiro and Rudin, who were the first to study its properties (see [9] and [8]). Golay's contribution was recently pointed out to us by Andrew Odlyzko. See [6], bottom of page 469.

2. GETTING STARTED: THE INITIAL QUESTION.

2. GETTING STARTED: THE INITIAL QUESTION. Many investigations begin with a question. In our case, we are looking at the terms of the Golay-Rudin-Shapiro sequence $\{a(n)\}$. This sequence can be defined recursively by the equations

$$\begin{aligned} a(2n) &= a(n), \\ a(2n+1) &= (-1)^n a(n), \quad n \geq 0, \\ a(0) &= 1. \end{aligned} \tag{1}$$

We know from work completed five years earlier [2] that the solution to this recurrence is

$$a(n) = (-1)^{e_0 e_1 + \cdots + e_{k-1} e_k}, \text{ where } n = \sum_{r=0}^{k} e_r 2^r, \quad e_r = 0 \text{ or } 1, \tag{2}$$

so this is clearly a sequence of ± 1's. The exponent on -1 in (2) counts the number of pairs of consecutive 1's in the binary representation of n. Thus if $n = 115_{10} = 1110011_2$, we have $a(115) = (-1)^3 = -1$.

The first eight terms of the sequence, starting with $n = 0$, are $1, 1, 1, -1, 1, 1, -1, 1$, and the obvious question is: does the number of $+1$'s exceed the number of -1's as we go out in the sequence? It's fine to ask this question, but now what? We might try rephrasing the question: "If we add up the terms, do the successive sums remain positive?" Being mathematicians, we make up some notation. Let

$$s(n) = \sum_{k=0}^{n} a(k), \quad n \geq 0. \tag{3}$$

The question now becomes: "Is $s(n) > 0$ for $n \geq 0$?"

To get some idea of what is going on, we do some computing. Working by hand and using the binary for n in (2), we readily find the values listed in Table 1. Since the answer to our question is "yes" for n up to 15, let's use a computer to extend Table 1 to, say, $n = 32{,}000$, and look at $s(n)$ over a larger range. We find when we do this that $s(n) > 0$ up to $n = 32{,}000$, so we begin to believe that the answer is "yes" for all n. Now what do we do?[2] Let's look at the long table more closely, and see if it suggests an obvious next question. The first thing we notice is a general growth in size of $s(n)$, with minor local variations. We can see this in the next part of the table (see Table 1A). Besides staying positive, the values of $s(n)$ roughly rise up to a peak of 15 (at $n = 42$) and then drop back down again, like a wave. Examining the long table to the end, we find that $s(n)$ cycles through four more such "waves," and that these waves seem to increase in "amplitude" and "wavelength."

TABLE 1

n	$a(n)$	$s(n)$	n	$a(n)$	$s(n)$
0	1	1	8	1	5
1	1	2	9	1	6
2	1	3	10	1	7
3	−1	2	11	−1	6
4	1	3	12	−1	5
5	1	4	13	−1	4
6	−1	3	14	1	5
7	1	4	15	−1	4

[2]This question is the hallmark of having temporarily run out of steam.

TABLE 1A

n	$s(n)$	n	$s(n)$	n	$s(n)$	n	$s(n)$	n	$s(n)$	n	$s(n)$
16	5	24	7	32	9	40	13	48	11	56	9
17	6	25	6	33	10	41	14	49	10	57	10
18	7	26	5	34	11	$\underline{42}$	$\underline{15}$	50	9	58	11
19	6	27	6	35	10	43	14	51	10	59	10
20	7	28	7	36	11	44	13	52	9	60	9
21	8	29	8	37	12	45	12	53	8	61	8
22	7	30	7	38	11	46	13	54	9	62	9
23	8	31	8	39	12	47	12	55	8	63	8

The obvious next question is: does this wavelike behavior continue? How do we turn this qualitative question into a precise mathematical question that we can actually work with?

Perhaps a first step would be to focus on one aspect of this wave, say its "wavelength." What does "wavelength" mean? In a strictly periodic wave, the wavelength is the distance between consecutive abscissae at which high points occur, for example. How shall we think about wavelength in the present context, where the wave is not periodic? The high points may still give us some indication, so let's look at the high points in the long table, and see what they tell us. (At this point the reader may want to make his or her own table; it may also be helpful to create a plot of the values $s(n)$, say, for n in the interval $[1, 64]$.)

As we look at the table, we notice a "strong" local maximum at various places. For example, $s(10) = 7$ is a clear local maximum. If we look further in the table, we notice an obvious sequence of these strong local maxima: $s(42) = 15$, $s(170) = 31$, and $s(682) = 63$. We see the beginning of a pattern. The s-values at these strong local maxima are 1 less than consecutive powers of 2, and each corresponding n-value is 4 times the previous one plus 2. Now we're getting somewhere! Next question: Does this pattern continue? We think it might, so we state our guess more formally.

Conjecture 1. The n-value for a strong local maximum is 2 more than 4 times the previous one; the s-values at these points are 1 less than successive powers of 2.

The first thing to do after making the conjecture is to test it. According to the conjecture, the next strong local maximum should occur at $n = 4 \cdot 682 + 2 = 2730$, and we should have $s(2730) = 2^7 - 1 = 127$. An examination of the table shows this is true, and this strengthens the conjecture. We now find ourselves being distracted from the original question by our conjecture, which is interesting in its own right. Let's indulge ourselves and pursue this pattern question, ignoring the original question for the time being. Experience shows that such side questions often connect back to the original question and give information that is important to the overall development.

To state a more precise conjecture, we need to develop a formula for the n-values at which these strong local maximum values seem to occur. Extending the sequence $n = 10, 42, 170, 682, 2730$ backwards, we find that $s(2) = 3$ is also a local maximum. If we set $M_0 = 2$, $M_1 = 10$, and in general, M_k equal to the n-value of the k-th local maximum, then the sequence $\{M_k\}$ is defined recursively by

$$M_{k+1} = 4M_k + 2, \quad M_0 = 2. \tag{4}$$

Using standard techniques we find that the solution to this recurrence is $M_k = 2(2^{2k+2} - 1)/3$, for $k \geq 0$. We can now make our previous conjecture more precise.

Conjecture 1′. The points $M_k = 2(2^{2k+2} - 1)/3$, $k \geq 0$, are the n-values of local maxima for $s(n)$, and $s(M_k) = 2^{k+2} - 1$.

Now what? Now that we have a conjecture about the high points, what can we say about the wavelength? If we take our definition of wavelength to be the difference between abscissae of successive strong local maxima, then the wavelength of the k-th wave cycle is $M_{k+1} - M_k$. Using (4) (still unproved), we find that $M_{k+2} - M_{k+1} = 4(M_{k+1} - M_k)$. Hence, our wavelengths increase by a factor of 4 from one wave to the next!

While staring at Conjecture 1′, it occurs to us that we should probably examine local minima, as well. It seems natural to think of a wave cycle beginning at a high point and ending at the next high point, so we decide to create a table (Table 2) of the absolute minima between high points, that is, in intervals of the form $[M_k, M_{k+1} - 1]$, $k \geq 0$. In successive intervals there appears to be a doubling of the number of n-values at which the minima occur. At this point we begin to wonder if this choice of interval will really lead to the simplest description of the behavior of the function. What would happen if we considered a different set of intervals? We notice that the last minimum in each interval of Table 2 occurs just before a power of 4. Putting this together with the fact that the wavelength increases by a factor of 4 from one wave to the next, we decide to make a list (Table 3) of the extrema in intervals of the form $[4^k, 4^{k+1} - 1]$, $k \geq 0$.

TABLE 2. n-VALUES FOR MINIMA IN $[M_k, M_{k+1} - 1]$

Interval	Minimum of $s(n)$	n-values at which minimum occurs
[2, 9]	2	3
[10, 41]	4	13, 15
[42, 169]	8	53, 55, 61, 63
[170, 681]	16	213, 215, 221, 223, 245, 247, 253, 255
[682, 2729]	32	853, 855, 861, 863, 885, 887, 893, 895, 981, 983, 989, 991, 1013, 1015, 1021, 1023

TABLE 3. n-VALUES FOR MINIMA AND MAXIMA IN $[4^k, 4^{k+1} - 1]$

Interval	Minimum of $s(n)$	n-values at which minimum occurs	Maximum of $s(n)$	n-values at which maximum occurs
[1, 3]	2	1, 3	3	2
[4, 15]	3	4, 6	7	10
[16, 63]	5	16, 26	15	42
[64, 255]	9	64, 106	31	170
[256, 1023]	17	256, 426	63	682
[1024, 4095]	33	1024, 1706	127	2730
[4906, 16383]	65	4096, 6826	255	10922

There is a surprising simplification here: there are now exactly two minima and one maximum in each power of 4 interval. Moreover, the sequence of n-values $6, 26, 106, 426, \ldots$, at which the second minimum occurs in each interval, satisfies the same recursion formula (4) that the numbers M_k did. This choice of interval has yielded gold!

Note the irrational element in the step we have just taken. There was no reason other than a desire for simplicity (or curiosity, or laziness?) for changing the interval. It turned out to be a good guess, but such a simplification may not always happen. In the present case, it seems quite remarkable that a simple shift of the interval has such a dramatic effect in cutting down the number of minima.

The first n-value of the pair giving the minimum in Table 3 is a power of 4. If we denote the second n-value of this pair by m_k, for $k \geq 1$ (starting with the second interval), we have the recursion $m_{k+1} = 4m_k + 2$, with $m_1 = 6$. Then we find easily that $m_k = (5 \cdot 4^k - 2)/3$, for $k \geq 1$. This leads to a more complete conjecture.

Conjecture 2. a) For $k \geq 1$, the minimum value of $s(n)$ in $[4^k, 4^{k+1} - 1]$ is $2^k + 1$, which occurs at just the points $n = 4^k$ and $n = m_k$. b) For $k \geq 0$, the maximum value of $s(n)$ in $[4^k, 4^{k+1} - 1]$ is $2^{k+2} - 1$, which occurs only at the point $n = M_k$.

Notice that if Conjecture 2 is correct, then it follows immediately that $s(n) > 0$ for $n \geq 0$. Notice also that we haven't yet proved *anything*!

How shall we go about trying to prove Conjecture 2? The fact that M_k and m_k satisfy the same recursion suggests that we try finding a formula for $s(4n + 2)$. Such a formula might be useful in proving the conjecture. Finding this formula turns out to be fairly simple, if we take a hint from (1) and first look for formulas for $s(2n)$ and $s(2n + 1)$. Using (1) to manipulate the sum which defines $s(2n)$ gives

$$s(2n) = \sum_{k=0}^{2n} a(k) = \sum_{k=0}^{n} a(2k) + \sum_{k=0}^{n-1} a(2k+1)$$
$$= \sum_{k=0}^{n} a(k) + \sum_{k=0}^{n-1} (-1)^k a(k) = s(n) + t(n-1),$$

where $t(n)$ is the new function defined by

$$t(n) = \sum_{k=0}^{n} (-1)^k a(k), \quad n \geq 0. \tag{5}$$

In the same way we can establish the following recursion formulas.

Lemma 1.

$$s(2n) = s(n) + t(n-1), \quad n \geq 1; \tag{6}$$
$$s(2n+1) = s(n) + t(n), \quad n \geq 0; \tag{7}$$
$$t(2n) = s(n) - t(n-1), \quad n \geq 1; \tag{8}$$
$$t(2n+1) = s(n) - t(n), \quad n \geq 0. \tag{9}$$

Uşing this lemma, we can work out a formula for $s(4n + 2)$. Replacing n by $2n + 1$ in (6) gives

$$s(4n+2) = s(2n+1) + t(2n) = s(n) + t(n) + s(n) - t(n-1)$$
$$= 2s(n) + (-1)^n a(n).$$

The lucky thing is that this recursion involves only the s-function, the t-function having dropped out. While we're at it, we give the formulas for $s(4n + d)$, where $d = 0, 1, 2, 3$ (there are similar formulas for $t(4n + d)$). The proofs are equally simple.

Lemma 2.

$$s(4n) = 2s(n) - a(n), \qquad n \geq 1; \tag{10}$$

$$s(4n+1) = s(4n+3) = 2s(n), \quad n \geq 0; \tag{11}$$

$$s(4n+2) = 2s(n) + (-1)^n a(n), n \geq 0. \tag{12}$$

It seems now that we might have enough relations to attempt an inductive proof of Conjecture 2. Since the assertions we have to prove concern the interval $[4^k, 4^{k+1} - 1]$, it makes sense to carry out the induction on the variable k. Let's look at the simpler statement in part b) of Conjecture 2: "For $k \geq 0$, the maximum value of $s(n)$ in $[4^k, 4^{k+1} - 1]$ is $2^{k+2} - 1$, which occurs only at the point $n = M_k$."

For $k = 0$ this assertion is true, since the maximum of $s(n)$ in $[1,3]$ is 3, which occurs only at $M_0 = 2$. Assume the assertion is true for the interval $I_k = [4^k, 4^{k+1} - 1]$ and try to prove it for I_{k+1}.

We first show that $2^{k+3} - 1$ is an upper bound for $s(n)$ in I_{k+1}. If n lies in $I_{k+1} = [4^{k+1}, 4^{k+2} - 1]$, then we can write $n = 4m + d$, for some m in $[4^k, 4^{k+1} - 1]$ and for some d in the set $\{0, 1, 2, 3\}$. Formulas (10)–(12) of Lemma 2 give:

$$s(n) = s(4m+d) \leq 2s(m) + 1 \leq 2(2^{k+2} - 1) + 1 = 2^{k+3} - 1, \tag{13}$$

which establishes the upper bound. Note that equality holds in (13) at most when $m = M_k$, by the induction assumption.

To finish the proof of b), we have to prove that M_{k+1} is the only place in I_{k+1} where the value $2^{k+3} - 1$ is actually achieved. This is really an "if and only if" statement. First we note from (4) and (12) that

$$s(M_{k+1}) = s(4M_k + 2) = 2s(M_k) + (-1)^{M_k} a(M_k). \tag{14}$$

Using (4) it is not hard to see that the binary expansion of M_k is $1010\ldots10$, with $k + 1$ occurrences of the pattern "10," an expansion that contains no consecutive 1's. Thus, $a(M_k) = +1$, and M_k is even, so the last term in (14) is $+1$, and the induction hypothesis leads to $s(M_{k+1}) = 2(2^{k+2} - 1) + 1 = 2^{k+3} - 1$.

Conversely, suppose that $s(n) = 2^{k+3} - 1$, for some n in I_{k+1}. We have to show that $n = M_{k+1}$. From the comment following (13) we get that $m = M_k$, so $n = 4M_k + d$, for some $d = 0$, 1, 2, or 3. Since $s(n)$ is odd, equation (11) shows that $d \neq 1$ or 3. If $d = 0$, formula (10) gives the contradiction

$$2^{k+3} - 1 = s(n) = s(4M_k) = 2s(M_k) - a(M_k) = 2(2^{k+2} - 1) - 1 = 2^{k+3} - 3.$$

Hence we must have $d = 2$, so $n = 4m + 2 = M_{k+1}$. This proves everything!

We leave it to the reader to prove part a)—the same technique works. So we have a theorem:

Theorem 1. *a) For $k \geq 1$, the minimum value for $s(n)$ in $[4^k, 4^{k+1} - 1]$ is $2^k + 1$, and this value occurs only at the points $n = 4^k$ and $n = m_k = (5 \cdot 4^k - 2)/3$. b) For $k \geq 0$, the maximum value for $s(n)$ in $[4^k, 4^{k+1} - 1]$ is $2^{k+2} - 1$, and this value occurs only at the point $n = M_k = \frac{2}{3}(2^{2k+2} - 1)$.*

Corollary. *The sum $s(n)$ is positive, for all $n \geq 1$.*

3. GOING FURTHER: A NEW QUESTION. Having proved more than we needed to settle the original question, we're hooked! Finding the answer to one question often suggests new questions. In our case the first new question was: what is the

asymptotic behavior of $s(n)$; that is, can we find an elementary function $f(n)$ so that the ratio $s(n)/f(n)$ stays bounded away from 0 and ∞, or even approaches a non-zero limit, as $n \to \infty$?

Finding such a function is not difficult, since Theorem 1 shows that $s(n)$ is roughly 2^k when n lies in the interval $I_k = [4^k, 4^{k+1} - 1]$. Hence, a reasonable choice is $f(n) = \sqrt{n}$. Looking at Theorem 1 more closely gives a good bit more. If n lies in the interval I_k, part a) shows that

$$\frac{s(n)}{\sqrt{n}} > \frac{2^k + 1}{\sqrt{4^{k+1}}} = \frac{1}{2} + \frac{1}{2^{k+1}} > \frac{1}{2},$$

while part b) gives that

$$\frac{s(n)}{\sqrt{n}} \le \frac{2^{k+2} - 1}{\sqrt{4^k}} = 4 - \frac{1}{2^k} < 4.$$

Thus, $\frac{1}{2} < s(n)/\sqrt{n} < 4$, for $n \ge 1$, and so $s(n)$ is roughly a constant times $\sqrt{n}$. This shows that $\sqrt{n}$ is the "right" order of magnitude of $s(n)$. However, this leaves open the question whether $s(n)/\sqrt{n}$ actually approaches a limit as $n \to \infty$. We can investigate this question using the sub-sequences $\{m_k\}$ and $\{M_k\}$. Using Theorem 1 we easily calculate that

$$\lim_{k\to\infty} \frac{s(m_k)}{\sqrt{m_k}} = \lim_{k\to\infty} \frac{(2^k + 1)\sqrt{3}}{\sqrt{5 \cdot 4^k - 2}} = \sqrt{\frac{3}{5}} = .7745\ldots$$

and

$$\lim_{k\to\infty} \frac{s(M_k)}{\sqrt{M_k}} = \lim_{k\to\infty} \frac{(2^{k+2} - 1)\sqrt{3}}{\sqrt{2(2^{2k+2} - 1)}} = \sqrt{6} = 2.4494\ldots.$$

Therefore $s(n)/\sqrt{n}$ certainly does not approach a limit as $n \to \infty$.

Having determined that the sequence has at least two limit points, two more questions occur to us: 1) how many limit points does the sequence have; and 2) what are the extremal limit points, i.e., the lim inf and lim sup of $\{s(n)/\sqrt{n}\}_{n=1}^{\infty}$? Such questions are often difficult to answer explicitly for the usual garden variety, pathological objects in real analysis. However, we have an intuition about the first question: since the denominator of $s(n)/\sqrt{n}$ is steadily increasing with n, while the numerator varies up and down by steps of 1, the difference between consecutive terms of the sequence is roughly

$$\frac{s(n+1)}{\sqrt{n+1}} - \frac{s(n)}{\sqrt{n}} \approx \frac{s(n+1) - s(n)}{\sqrt{n}} = \frac{\pm 1}{\sqrt{n}},$$

so the difference tends to 0 as n increases. Thus, an increasing number of smaller and smaller steps will be required to pass from the neighborhood of one limit point to the other. It is reasonable to guess, then, that *every* point of the interval $[\sqrt{3/5}, \sqrt{6}]$ is a limit point of $\{s(n)/\sqrt{n}\}_{n=1}^{\infty}$. This is true, in fact, and it isn't hard to turn our intuition into a formal proof.

Theorem 2. *Every point of the interval* $[\sqrt{3/5}, \sqrt{6}]$ *is a limit point of the sequence* $\{s(n)/\sqrt{n}\}_{n=1}^{\infty}$.

Proof: Let $\sqrt{3/5} < \xi < \sqrt{6}$, and suppose $\delta > 0$ and $N > 0$ are given. The assertion is: there is some $n > N$ for which

$$\left|\frac{s(n)}{\sqrt{n}} - \xi\right| < \delta. \tag{15}$$

Choose $m_2 > m_1 > N$ so that both $s(m_1)/\sqrt{m_1} < \xi < s(m_2)/\sqrt{m_2}$ and $\sqrt{n} > 1/\delta$ for $n \geq m_1$. Then, as we now show, (15) will be satisfied for some n between m_1 and m_2. Suppose to the contrary that there is no n with $m_1 \leq n \leq m_2$ for which (15) is satisfied, i.e., for which $\xi - \delta < s(n)/\sqrt{n} < \xi + \delta$. Then there must be a largest n_1 in $[m_1, m_2)$ for which $s(n_1)/\sqrt{n_1} \leq \xi - \delta$. This implies $s(n_1+1)/\sqrt{n_1+1} \geq \xi + \delta$, and therefore, $s(n_1+1)/\sqrt{n_1+1} - s(n_1)/\sqrt{n_1} \geq 2\delta$. On the other hand,

$$\frac{s(n_1+1)}{\sqrt{n_1+1}} - \frac{s(n_1)}{\sqrt{n_1}} \leq \frac{s(n_1+1) - s(n_1)}{\sqrt{n_1}} \leq \frac{1}{\sqrt{n_1}} < \delta.$$

The contradiction $2\delta < \delta$ shows our supposition to be false, i.e., there does exist an integer $n > N$ for which (15) is true. ∎

This answers question 1). To get some idea about question 2), we go back to the computer and compute the terms $s(n)/\sqrt{n}$ to $n = 32{,}000$. (Again, the reader may want to create a table or a plot of $s(n)/\sqrt{n}$ to follow along in this discussion.) We see the high points at $n = 10, 42, 170$ and low points at $n = 15, 26, 106$. The maximum of $s(n)/\sqrt{n}$ in I_k occurs at exactly the same point $n = M_k$ that it does for $s(n)$. Except for $n = 15$—an unruly initial exception—the minimum in I_k occurs at $n = m_k$ and is unique.

If extremal points for $s(n)$ remain extremal for the quotient $s(n)/\sqrt{n}$, it is plausible to think that the limit points $\sqrt{3/5}$ and $\sqrt{6}$ might be the lim inf and lim sup of the quotient sequence. Moreover, as is easily shown, the terms $s(m_k)/\sqrt{m_k}$ decrease monotonically to $\sqrt{3/5}$ and the terms $s(M_k)/\sqrt{M_k}$ increase monotonically to $\sqrt{6}$. Hence it is also plausible that $\sqrt{3/5}$ and $\sqrt{6}$ might be upper and lower bounds as well. Polya encourages us to be bold,[3] so we take the leap:

Conjecture 3. For $n \geq 1$, $\sqrt{3/5} < s(n)/\sqrt{n} < \sqrt{6}$.

If this is true, then it certainly follows that $\sqrt{3/5}$ and $\sqrt{6}$ are the lim inf and lim sup, respectively, since we already know they are limit points. Hence we can focus all of our attention on Conjecture 3.

4. THE UPPER BOUND. We decide to focus on the upper bound first, since the maxima of $s(n)/\sqrt{n}$ seem to occur for the same n-values that they do for $s(n)$. In taking this route, we're letting ourselves be guided by the sense of internal or hidden logic that the subject seems to have.

What should we try? Suppose we focus on the interval $I_k = [4^k, 4^{k+1} - 1]$, as before. We would like to show that $s(n)/\sqrt{n} < \sqrt{6}$ for n in I_k, and we know that

[3] See George Polya's lecture in "Let Us Teach Guessing," an MAA video.

$s(M_k)/\sqrt{M_k} < \sqrt{6}$. It is natural to try to show that

$$(16) \qquad s(n)/\sqrt{n} \le s(M_k)/\sqrt{M_k}, \text{ for } n \text{ in } I_k.$$

We know that $s(n) \le s(M_k)$ on the whole interval, but we can only say that $\sqrt{n} \ge \sqrt{M_k}$ when $n \ge M_k$; hence the inequality in (16) is true in the sub-interval $[M_k, 4^{k+1} - 1]$. This is a start.

This is a common situation in mathematics. Working on a problem can be very much like putting together a jigsaw puzzle. You first try to find pieces you can fit together, so as to form islands of connections. Having found some of these islands, you then want to see how they fit together to solve the larger puzzle. The trouble with a mathematical puzzle is that you don't always know what all the pieces *are*; sometimes you have to create them. Sometimes, they don't all fit together to make the picture you want.

To create the next part of the argument we have to try imagining a new piece of the puzzle, something that will give us another way to look at the inequality (16). We have proved the inequality in the interval $[M_k, 4^{k+1} - 1]$, and we want to prove it in the interval $[4^k, M_k]$. We need some formulas to work with.

What do we know? We have the formulas (10)–(12), which relate values of s in I_k to values in I_{k-1}. It might be worth trying to use these as part of an induction proof. We try formula (10) first:

$$\frac{s(4n)}{\sqrt{4n}} = \frac{2s(n) - a(n)}{\sqrt{4n}} \le \frac{2s(n) + 1}{2\sqrt{n}} = \frac{s(n)}{\sqrt{n}} + \frac{1}{2\sqrt{n}}.$$

This almost works, except for that annoying term $1/2\sqrt{n}$. On the other hand, we find that formula (11) really does work, giving us $s(4n+1)/\sqrt{4n+1} < \sqrt{6}$ and $s(4n+3)/\sqrt{4n+3} < \sqrt{6}$, if $s(n)/\sqrt{n} < \sqrt{6}$. We get the desired inequality for *odd* values of n in I_k if we know it for *all* values in I_{k-1}. Close, but not good enough for an induction proof.

It seems as if these formulas don't quite give us what we want. We decide to go back to the tables and look for other patterns that might give a clue to some useful relationships. First we notice the similarity between the first 8 values of $s(n)$ in Tables 1 and 1A. We list them side by side:

Table 1 ($0 \le n \le 7$)	1	2	3	2	3	4	3	4
Table 1A ($16 \le n \le 23$)	5	6	7	6	7	8	7	8

The difference between corresponding numbers in the two rows is always 4. The next eight values from each table are:

Table 1 ($8 \le n \le 15$)	5	6	7	6	5	4	5	4
Table 1A ($24 \le n \le 31$)	7	6	5	6	7	8	7	8

Now the *sum* of corresponding numbers is always 12. We summarize these patterns in equation form:

$$s(n+16) = s(n) + 4, \qquad 0 \le n \le 7;$$

$$s(n+16) = -s(n) + 12, \quad 8 \le n \le 15.$$

When we consider the last four columns in Table 1A we find similar patterns:

$$s(n+32) = s(n) + 8, \qquad 0 \le n \le 15;$$

$$s(n+32) = -s(n) + 16, \quad 16 \le n \le 31.$$

These four equations show that all the values of $s(n)$ for n in $[16, 63]$ are obtainable from the values in $[0, 15]$. Looking further in the table we see that these

elaborate patterns continue, and we are able to guess the general form of the last term in each of the four equations.

Lemma 3.

$$s(n+2^{2k}) = s(n) + 2^k, \qquad 0 \le n \le 2^{2k-1}-1,\ k \ge 1; \tag{17}$$

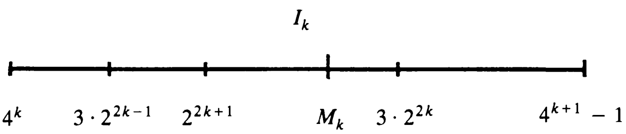
$$s(n+2^{2k}) = -s(n) + 3\cdot 2^k, \qquad 2^{2k-1} \le n \le 2^{2k}-1,\ k \ge 1; \tag{18}$$

$$s(n+2^{2k+1}) = s(n) + 2^{k+1}, \qquad 0 \le n \le 2^{2k}-1,\ k \ge 0; \tag{19}$$

$$s(n+2^{2k+1}) = -s(n) + 2^{k+2}, \qquad 2^{2k} \le n \le 2^{2k+1}-1,\ k \ge 0. \tag{20}$$

The proof of the formulas, once we have found them, is not hard. The idea is to prove them by induction on n, for a fixed k (see [**3**, Satz 5]).

These formulas show that the values of the sequence $s(n)$, for n in $I_k = [4^k, 4^{k+1}-1]$, can be generated from the values in $[0, 4^k - 1]$. In this process the interval I_k is divided into four pieces, the last two of which are twice as long as the first two. The integer $M_k = 2(2^{2k+2}-1)/3$ is contained in the third of these subintervals.



This method of generating the sequence gives us some hope that we can prove the inequalities we want to. Formula (19) catches our eye first because $M_k = 2 + 2^3 + \cdots + 2^{2k+1}$ is a sum of odd powers of 2, so that $M_{k-1} + 2^{2k+1} = M_k$ (see the discussion following (14)). Hence we see that if $n + 2^{2k+1} \le M_k$, then $n \le M_{k-1}$. This looks promising, but there is a restriction on the values of n for which formula (19) holds: n must lie between 0 and $2^{2k}-1$. Thus we can make use of (19) only for values of $n + 2^{2k+1}$ lying in the interval

$$2^{2k+1} \le n + 2^{2k+1} \le 2^{2k} - 1 + 2^{2k+1} = 3 \cdot 2^{2k} - 1.$$

Since M_k is definitely in this interval, we can use formula (19) in the subinterval $[2^{2k+1}, M_k]$. We find that

$$\frac{s(n+2^{2k+1})}{\sqrt{n+2^{2k+1}}} = \frac{s(n)}{\sqrt{n}}\,\frac{\sqrt{n}}{\sqrt{n+2^{2k+1}}} + \frac{2^{k+1}}{\sqrt{n+2^{2k+1}}} < \frac{\sqrt{6n}+2^{k+1}}{\sqrt{n+2^{2k+1}}} \tag{21}$$

for $1 \le n \le M_{k-1}$, assuming that the inequality $s(n)/\sqrt{n} < \sqrt{6}$ is known for $n \le M_{k-1}$.[4] We get the inequality we want as long as

$$\frac{\sqrt{6n}+2^{k+1}}{\sqrt{n+2^{2k+1}}} < \sqrt{6}, \tag{22}$$

i.e., as long as $\sqrt{6n} + 2^{k+1} < \sqrt{6n + 3\cdot 2^{2k+2}}$. The latter inequality is equivalent to $n < 2^{2k+1}/3$, which is true in the interval we're considering ($n \le M_{k-1}$).

[4]Note that we have to assume $n \ge 1$ since we divide by $\sqrt{n}$ in (25).

We now have part of an induction step. This would prove $s(n)/\sqrt{n} < \sqrt{6}$ in $[2^{2k+1}+1, M_k]$.

Note that this analysis is already more complicated than the analysis in Section 2. This is to be expected, since the new question involves a ratio of functions.

The only part of the interval I_k we have yet to consider is $[2^{2k}, 2^{2k+1}]$. We might be able to use (17) in a similar way to handle this subinterval, but this formula would only be applicable for $n + 2^{2k} \le 3 \cdot 2^{2k-1} - 1$, leaving out the subinterval $[3 \cdot 2^{2k-1}, 2^{2k+1}]$. However, we also notice from our tables that $s(n)$ appears to be bounded by 2^{k+1} in $[2^{2k}, 2^{2k+1} - 1]$, which would give us the estimate

$$\frac{s(n)}{\sqrt{n}} \le \frac{2^{k+1}}{2^k} = 2 < \sqrt{6}, \tag{23}$$

for $k \ge 0$. The one omitted value in (23), $n = 2^{2k+1}$, can be checked using the fact that $s(2^{2k+1}) = 2^{k+1} + 1$:

$$\frac{s(2^{2k+1})}{\sqrt{2^{2k+1}}} = \frac{2^{k+1}+1}{\sqrt{2}\cdot 2^k} < \sqrt{2} + 1 < \sqrt{6}, \tag{24}$$

for $k \ge 0$. So all we have to do to make this work is prove that $s(n) \le 2^{k+1}$, for n in $[2^{2k}, 2^{2k+1} - 1]$. The proof of this last assertion is very similar to the proof of Theorem 1b), as long as we also specify the places where $s(n)$ takes on the value 2^{k+1}. This gives the following statement.

Lemma 4. *For $2^{2k} \le n \le 2^{2k+1} - 1$ we have $s(n) \le 2^{k+1}$, with equality if and only if $n = 2^{2k+1} - 1 - \sum_{r=0}^{k-1} \varepsilon_r 2^{2k+1}$, where $\varepsilon_r = 0$ or $1, k \ge 0$.*

Again we leave the straightforward induction proof to the reader (see [**3**, Satz 9]). The condition for equality in this lemma comes from analyzing the binary representations of the n-values at which equality holds.

We have all the pieces now!

Theorem 3. *For $n \ge 1$, $s(n)/\sqrt{n} < \sqrt{6}$.*

Proof: We just sketch the outline, since we have already given most of the details above. We focus on values of n in the interval $[4^k, 4^{k+1} - 1]$. In the subinterval $[M_k, 4^{k+1} - 1]$, $k \ge 0$, we use the argument in the paragraph containing (16). In the interval $[2^{2k}, 2^{2k+1}]$, $k \ge 0$, we use Lemma 4, (23) and (24). To prove the inequality in the remaining interval $[2^{2k+1}+1, M_k]$, for $k \ge 1$, we use induction on k. The assertion is true for $k = 1$, since the interval in question is $[9, 10]$, and the inequality can be checked directly. If the assertion is true for $k - 1$, then (19), (21) and (22) show that it also holds for k. It is easy to check that this covers all the integers, and the proof is complete.

There is no way of telling whether this proof for the upper bound is the simplest proof. Perhaps the reader can find a better one.

5. THE LOWER BOUND. What about the lower bound? Will a similar approach establish that $s(n)/\sqrt{n} > \sqrt{3/5}$ for $n \ge 1$? Looking over what we've done, we observe that we used three different approaches in three different subintervals of I_k to establish the upper bound. In tackling the lower bound, we wonder if a more

unified approach will work, one that just uses the formulas of Lemma 3 on each of the four subintervals of I_k determined by that lemma. We denote these four subintervals by I, II, III, IV.

On interval I, a simple argument using the lower bound $s(n) \geq 1$ in (17) gives

$$\frac{s(n+2^{2k})}{\sqrt{n+2^{2k}}} = \frac{s(n)+2^k}{\sqrt{n+2^{2k}}} \geq \frac{1+2^k}{\sqrt{3\cdot 2^{2k-1}-1}} > \frac{1+2^k}{\sqrt{3\cdot 2^{2k-1}}} = \sqrt{\frac{2}{3}}\,\frac{1+2^k}{2^k},$$

and the right hand side is obviously $> \sqrt{3/5}$, for $k \geq 1$. The same argument works in interval III, using (19), for $k \geq 0$. A similar argument using (20) gives the lower bound we want in interval IV (hint: use the upper bound from Lemma 4).

We're almost home. The last subinterval II is the trickiest one, since it contains the minimum point m_k. As it happens, we now run out of luck! It doesn't miss by much, but formula (18) is apparently not strong enough to allow us to prove what we believe to be true in the whole of interval II. (The reader may enjoy performing the calculations.) Unfortunately, we have found no way to finish the proof of the lower bound by this method, i.e., by piecing together inequalities on subintervals of I_k, even though it shows that the inequality we want is true for most integers. The interval idea leads to a dead end. R.I.P.[5]

This is a good example of a proof that fails by the slimmest of margins. However, this was only our first attempt to prove the lower bound. We tried to imitate what worked nicely for the upper bound, and got stuck. It might be that the interval method didn't work to establish the lower bound because the method was simply too crude. We need to go back and look at the problem again, if possible, from a different point of view.

How else can we look at the problem of getting a lower bound for $s(n)/\sqrt{n}$? As we look at the graph of $s(n)$, we see that there are only a finite number of places n where $s(n)$ has a fixed value k, since the values of $s(n)$ go to infinity with n (by Theorem 1a). When $s(n) = k$, for a fixed k, the ratio $s(n)/\sqrt{n}$ will be smallest when n is largest. This leads us to the following idea: let's focus on the last (largest) value of n for which a given integer k appears as $s(n)$ in the sequence $\{s(n)\}_{n=0}^{\infty}$; call it $\omega(k)$. If we could prove the inequality

$$k/\sqrt{\omega(k)} > \sqrt{3/5}\,, \text{ for } k \geq 1, \tag{25}$$

then the lower bound would follow for $s(n)/\sqrt{n}$, for any n, because taking $k = s(n)$ would give $s(n)/\sqrt{n} \geq s(n)/\sqrt{\omega(s(n))} > \sqrt{3/5}$. Thus the subsequence $\{k/\sqrt{\omega(k)},\ k \geq 1\}$ of $\{s(n)/\sqrt{n},\ n \geq 1\}$ is the key subsequence to consider in looking for the best lower bound for $s(n)/\sqrt{n}$. Not knowing what else to do, we set off to see if we can prove (25). First, we state the definition of ω formally.

Definition. For a given $k \geq 1$, let $\omega(k)$ be the largest value of n for which $s(n) = k$.

The next thing to do is to go back to our tables and determine $\omega(k)$ for the values of k up to 255. The beginning of the table is given in Table 4. In the full table we are surprised to find that $\omega(k)$ satisfies recursion formulas much like those satisfied by $s(n)$, but with some interesting wrinkles.

[5] Ending symbol for a proof that didn't work.

TABLE 4

k	1	2	3	4	5	6	7	8	9	10
$\omega(k)$	0	3	6	15	26	27	30	63	106	107

Lemma 5.

$$\omega(2n) = 4\omega(n) + 3, \qquad n \geq 1. \tag{26}$$

$$\omega(2n+1) = 4\omega(n+1) + 2, \quad n \geq 2, n+1 \neq 2^r, r \geq 2. \tag{27}$$

Proof: The proof of (26) is not hard. Note first that $s(n)$ and n have opposite parity, so that $\omega(2n)$ must be odd. Hence we have either that $\omega(2n) = 4m + 1$ or $\omega(2n) = 4m + 3$, for some $m \geq 0$. The first case is impossible, because by (11), $s(4m+1) = s(4m+3) = 2n$, so $4m + 1$ cannot be the largest argument of s to give $2n$. Thus $\omega(2n) = 4m + 3$. Then (11) implies that $2n = s(4m+3) = 2s(m)$, so that $s(m) = n$. If there were an $m_1 > m$ with $s(m_1) = n$, then $s(4m_1 + 3) = 2n$ and $4m_1 + 3 > 4m + 3 = \omega(2n)$ would contradict the definition of $\omega(2n)$. Thus $\omega(n) = m$, and $\omega(2n) = 4\omega(n) + 3$.

The proof of (27) is much trickier. To see how to approach the proof, let's first note one consequence of the formula. If (27) is true, then certainly

$$s(4\omega(n+1)+2) = s(\omega(2n+1)) = 2n+1.$$

How might we prove just this much? Formula (12) gives

$$\begin{aligned} s(4\omega(n+1)+2) &= 2s(\omega(n+1)) + (-1)^{\omega(n+1)}a(\omega(n+1)) \\ &= 2(n+1) + (-1)^n a(\omega(n+1)), \end{aligned}$$

and so $s(4\omega(n+1)+2) = 2n+1$ if and only if $a(\omega(n+1)) = (-1)^{n+1}$ (when $n+1$ is not a power of 2). This shows that to prove (27) we must consider the formula

$$a(\omega(n)) = (-1)^n, \qquad n \geq 3, n \neq 2^r, r \geq 2. \tag{28}$$

Since induction has worked so often before, it is worth trying to prove (28) by induction as well. This is what we do now. Formula (28) holds for $n = 3$, since $a(\omega(3)) = a(6) = -1 = (-1)^3$. Assume that (28) has been proved for all the integers m for which $3 \leq m < 2n+1$, for some $n \geq 2$. We proceed to prove it for $2n+1$ and $2n+2$. There are two cases to consider, because of the excluded values in (28).

Case 1: Suppose that $2n+2 \neq 2^r$, for any $r \geq 3$. Since we have already proved (26), we can use that formula and the defining formulas (1) for $a(n)$ to compute $a(\omega(2n+2))$:

$$\begin{aligned} a(\omega(2n+2)) &= a(4\omega(n+1)+3) = -a(2\omega(n+1)+1) \\ &= (-1)^{1+\omega(n+1)}a(\omega(n+1)) \\ &= (-1)^{1+n}(-1)^{n+1} = (-1)^{2n+2}; \end{aligned}$$

this computation uses the fact that the parities of $\omega(n+1)$ and $n+1$ are opposite, and the induction assumption ($n+1 < 2n+1$ and $n+1$ is not a power of 2). This proves (28) for $2n+2$. Before considering (28) for $2n+1$, we need a

formula for $\omega(2n+1)$. From what we have just shown it is easy to find a good candidate for $\omega(2n+1)$, since

$$s(\omega(2n+2)-1)=s(\omega(2n+2))-a(\omega(2n+2))=(2n+2)-1=2n+1.$$

Thus we might guess that $\omega(2n+1)=\omega(2n+2)-1$. If there were an $m>\omega(2n+2)-1$ for which $s(m)=2n+1$, then because the sequence $\{s(m), m\geq 0\}$ goes to infinity by steps of ± 1, there would have to be an integer $m'>m$ for which $s(m')=2n+2$. But then $m'\geq m+1>\omega(2n+2)$ would give a contradiction. Hence our guess was correct, and $\omega(2n+1)=\omega(2n+2)-1=4\omega(n+1)+2$. This proves (27), since in this case $n+1$ is not equal to a power of 2.

Now (28) follows for the value $2n+1$, since

$$\begin{aligned} a(\omega(2n+1)) &= a(4\omega(n+1)+2)=a(2\omega(n+1)+1)\\ &= (-1)^{\omega(n+1)}a(\omega(n+1))\\ &= (-1)^{n}(-1)^{n+1}=(-1)^{2n+1}. \end{aligned}$$

Case 2: If $2n+2=2^r$, for some $r\geq 3$, then to complete the induction we have to prove (28) only for the value $2n+1=2^r-1$. Here we need the fact that $\omega(2^r-1)=2^{2r-1}-2$. To see this, first note that

$$s(2^{2r-1}-2)=s(2^{2r-1}-1)-a(2^{2r-1}-1)=2^r-1$$

by Lemma 4 and the fact that there are $2r-2$ pairs of consecutive 1's in the binary expansion of $2^{2r-1}-1$. Furthermore, it can be proved by induction that $s(m)\geq 2^r$ for $2^{2r-1}\leq m\leq 2^{2r}-1$ (with equality if and only if $m=2^{2r}-1-\sum_{i=0}^{r-2}\varepsilon_i 2^{2i+1}$, where $\varepsilon_i=0$ or 1). This, together with Theorem 1a (take $k\geq r$), shows that $s(m)\geq 2^r$ when $m>2^{2r-1}-2$, and hence that $\omega(2^r-1)=2^{2r-1}-2$, as claimed.

It follows that $a(\omega(2n+1))=a(\omega(2^r-1))=a(2^{2r-1}-2)=(-1)^{2r-3}=(-1)^{2n+1}$, and this completes the proof of (28). With (28) we have also completely proved (27) as well. ∎

Looking back over this proof, we see that we were led to (28) by considering possible consequences of (27), but then proving (28) gave us a complete proof of (27) as a by-product: formula (27) is implied by (28) at the value $2n+2$. Actually, if we think of the induction proof as an argument that proceeds step-by-step through the positive integers, then the two formulas (27) and (28) are really *intertwined*, since (28) at $2n+2$ is used to establish (27), which is used in turn to prove (28) at $2n+1$. It is surprising that such intricate arguments are required to establish fairly simple recursion formulas.

After we had found the recursion formulas in Lemma 5, it seemed we were no closer to a proof of (25). However, we started to look for more patterns in the table by taking differences between consecutive values of $\omega(n)$, one of the standard ways of spotting possible formulas. Taking differences of the first 25 terms of the ω sequence gives:

n	1	2	3	4	5	6	7	8	9	10	11	12
$\omega(n+1)-\omega(n)$	3	3	9	11	1	3	33	43	1	3	1	11

n	13	14	15	16	17	18	19	20	21	22	23	24
$\omega(n+1)-\omega(n)$	1	3	129	171	1	3	1	11	1	3	1	43

What strange numbers! For long stretches the difference is 1 at odd integers, and then it skyrockets. At even integers n, the difference takes on the values $3, 11, 3, 43$, except at powers of 2, where it also suddenly increases. Powers of 2! Suddenly we see that the difference depends only on the power of 2 dividing n, except when n is 1 less than a power of 2, a wrinkle that fits with the recursion formulas in Lemma 5. We also see that the values $1, 3, 11, 43, 171$ satisfy a recursion: each value is 4 times the preceding value minus 1. When we solve the recursion for these values and investigate the wrinkle more closely, we find the following remarkable formulas.

Lemma 6. *a) If* $n = 2^{\alpha}(2m+1)$, *for* $m \geq 0$, $\alpha \geq 0$, *then* $\omega(n+1) - \omega(n) = (2^{2\alpha+1}+1)/3$, *unless* $\alpha = 0$ *and* $n = 2^s - 1$, $s \geq 1$. *b) If* $\alpha = 0$ *and* $n = 2^s - 1$, $s \geq 1$, *then* $\omega(n+1) - \omega(n) = 2^{2s-1} + 1$.

We omit the details of the proof, and note only that part a) can be proved by induction on α, using (26), (27), and the special values $\omega(2^s - 1) = 2^{2s-1} - 2$ and $\omega(2^{s+2} - 2) = 2^{2s+3} - 5$.

Once we have a formula for the difference $\omega(n+1) - \omega(n)$, we are close to finding a formula for $\omega(k)$, since $\sum_{n=1}^{k-1}\{\omega(n+1) - \omega(n)\} = \omega(k) - \omega(1) = \omega(k)$. Summing up the expressions in Lemma 6 leads to the following explicit formula.

Theorem 4. *If* $2^r \leq k \leq 2^{r+1} - 1$, $r \geq 0$, *then*

$$\omega(k) = k - 1 + \frac{1}{3}(2^{2r+1} - 2) + 2\sum_{i=0}^{r-1}\left[\frac{k-1}{2^{i+1}}\right]2^{2i}.$$

We leave the somewhat technical details of the proof to the reader. This formula follows directly from Lemma 6, but may also be proved by a straightforward induction proof (on k) using only Lemma 5 and the fact that $\omega(2^r - 1) = 2^{2r-1} - 2$, $r \geq 0$. (See [**3**, Satz 1].)

Will this formula give us the lower bound we want?

Theorem 5. *For* $k \geq 1$, $k/\sqrt{\omega(k)} > \sqrt{3/5}$.

Proof: Assume that $2^r \leq k \leq 2^{r+1} - 1$, $r \geq 0$. By the formula for $\omega(k)$ we have

$$\begin{aligned} 3\omega(k) &= 3k - 3 + 2^{2r+1} - 2 + 6\sum_{i=0}^{r-1}\left[\frac{k-1}{2^{i+1}}\right]2^{2i} \\ &\leq 3k - 3 + 2^{2r+1} - 2 + 3(k-1)\sum_{i=0}^{r-1}2^i \\ &= 3k - 3 + 2^{2r+1} - 2 + 3(k-1)(2^r - 1) \\ &< 3k - 3 + 2k^2 - 2 + 3(k-1)k = 5k^2 - 5 < 5k^2 \end{aligned}$$

and the inequality of the theorem follows immediately. It worked!

Corollary. *For* $n \geq 1$, $s(n)/\sqrt{n} > \sqrt{3/5}$.

To summarize, we may combine Theorems 2, 3, and 5 in the following explicit result.

Theorem 6. *For* $n \geq 1$ *we have* $\sqrt{3/5} < s(n)/\sqrt{n} < \sqrt{6}$, *and the sequence* $\{s(n)/\sqrt{n},\ n \geq 1\}$ *is dense in the interval* $[\sqrt{3/5}, \sqrt{6}]$. *In particular,*

$$\limsup_{n\to\infty} \frac{s(n)}{\sqrt{n}} = \sqrt{6} \qquad \textit{and} \qquad \liminf_{n\to\infty} \frac{s(n)}{\sqrt{n}} = \sqrt{3/5}.$$

Looking back over our development, we find that our original purposes, which were to introduce the reader to the Golay-Rudin-Shapiro sequence, and to illustrate how mathematicians are led by their questions, have been realized. Sometimes, proofs of conjectures are constructed without difficulty and work more or less on the first attempt, as in Section 2.[6] On the other hand, this was not the case in the investigation of the lower bound, where a leap was required into a whole new investigation to get past a barrier in our first attempt at a proof. In the process we broke through into an area that is interesting in its own right, as is evidenced by the mysterious and elegant properties of the ω-function.

Since our main purpose has been to retrace the development of questions and ideas, we have given priority to these questions over full details of sometimes technical proofs. We hope the reader will find it an interesting challenge to fill in these details, or to read them in **[3]**. The reader can also obtain an expanded version of this paper by writing to the authors.

REFERENCES

1. J.-P. Allouche and J. Shallit, The ring of k-regular sequences, *J. Theoretical Computer Science* 98 (1992), 163–197.
2. J. Brillhart and L. Carlitz, Note on the Shapiro polynomials, *Proc. Amer. Math. Soc.* 25 (1970), 114–118.
3. J. Brillhart and P. Morton, Über Summen von Rudin-Shapiroschen Koeffizienten, *Illinois J. Math.* 22 (1978), 126–148.
4. J. Brillhart, P. Erdös and P. Morton, On Sums of Rudin-Shapiro Coefficients, II, *Pacific J. Math.* 107 (1983), 39–69.
5. M. Dekking, M. Mendes France, A. J. van der Poorten, Folds! I, II, III, *Math. Intelligencer* 4 (1982), 130–138, 173–181, 190–195.
6. M. J. E. Golay, Static multislit spectrometry and its application to the panoramic display of infrared spectra, *J. Optical Soc. America* 41 (1951), 468–472.
7. P. Morton and W. Mourant, Paper folding, digit patterns, and groups of arithmetic fractals, *Proc. London Math. Soc.* (3) 59 (1989), 253–293.
8. W. Rudin, Some theorems on Fourier coefficients, *Proc. Amer. Math. Soc.* 10 (1959), 855–859.
9. H. S. Shapiro, Extremal problems for polynomials and power series, Master's thesis, M.I.T., Cambridge, Mass., 1951.

Department of Mathematics
University of Arizona
Tucson, AZ 85721
jdb@math.arizona.edu

Department of Mathematics
Wellesley College
Wellesley, MA 02181
pmorton@wellesley.edu

[6]But recall the detailed analysis that led to precise conjectures before any proof had been attempted.

A Mathematician Catches a Baseball

Edward Aboufadel

1. INTRODUCTION. In the game of baseball, what strategy does an outfielder employ to catch a fly ball? Recently, Michael McBeath and Dennis Shaffer, who are psychologists, and Mary Kaiser, a researcher at NASA, proposed a new model to explain how this task is accomplished [1]. The model, called the linear optical trajectory (LOT) model, was developed and tested empirically by the three researchers, and it received national attention during the 1995 baseball season [4]. In this paper, seeking to clarify what is written in [1] and [4], we develop equations relating the motion of a fly ball to the motion of an outfielder utilizing the LOT strategy. In the process, we provide a mathematical foundation on which the LOT model can rest.

To begin, let H be home plate, B the position of the ball, and F the position of the fielder at any point in time; see Figure 1. Define B^* to be the projection of B onto the playing field, so that H, F, and B^* are co-planar. There is a well-defined point I^*, which is the point of intersection of the line B^*F and the unique perpendicular to B^*F through H. There is another well-defined point I, which lies on the line BF, directly above I^*. The point I is the fielder's image of the ball.

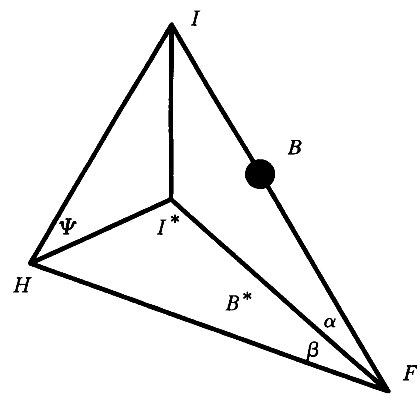


Figure 1

There are three important angles defined in the right pyramid $HFII^*$: the vertical optical angle $\alpha = \angle B^*FB$, the horizontal optical angle $\beta = \angle B^*FH$, and the optical trajectory projection angle $\Psi = \angle I^*HI$. We then have:

The LOT model hypothesis. *The strategy that a fielder uses to catch a fly ball is to follow a path that will keep the optical trajectory projection angle Ψ constant; this is equivalent to keeping the ratio* $(\tan \alpha)/(\tan \beta)$ *constant.*

In discussing the LOT model, McBeath and his colleagues write, "The LOT strategy discerns optical acceleration as optical curvature, a feature that observers are very good at discriminating," and, "If you're running along a path that doesn't allow the ball to curve down, then in a sense you are guaranteed to catch it." The LOT model apparently also applies to other situations, such as the pursuit of mates and prey by certain fish and houseflies. McBeath et al. also write, "[The LOT model] keeps the image of the ball continuously ascending in a straight line throughout the trajectory." [**1, 2**] This last statement can be misleading to the casual reader who assumes it means that the trajectory of I must be linear. In Section 4, we clarify this statement and others that have been made about the model.

2. THE FIELDER, HIS PREY, AND THE IMAGE OF HIS PREY. If a fielder uses the LOT model, is his path uniquely determined by the path of the baseball? In this section, we develop equations relating the positions of the fielder, the ball, and the image of the ball, as we seek an answer to this question. Our analysis is in three-dimensional space, with the playing field represented by the xy-plane. Without loss of generality, let home plate H be the origin, and identify the first and third base lines with the x-axis and the y-axis, respectively, so that a fair ball is one that lands in the first quadrant of the plane. The coordinates of our three relevant points are $F = (x_f, y_f, z_f)$, $B = (x_b, y_b, z_b)$, and $I = (x_i, y_i, z_i)$. All nine coordinates are functions of time t (with $t = 0$ representing the moment that the ball is hit by the batter), and $z_f = 0$ at all points in time.

We define two other functions of time:

$$p = \frac{y_i}{x_i} \quad \text{and} \quad q = \frac{z_i}{x_i}. \tag{1}$$

If the trajectory of I is linear, then p and q would be constant. However, in the LOT model, this is not necessary. Instead, we have the following lemma:

Lemma 2.1. *If the LOT model is valid (i.e., if Ψ is constant), then $q^2/(1+p^2)$ is constant.*

Proof of Lemma 2.1: If we consider Figure 1, we see that

$$\tan^2 \Psi = \frac{|II^*|^2}{|HI^*|^2} = \frac{z_i^2}{x_i^2 + y_i^2} = \frac{q^2}{1+p^2}. \quad \blacksquare \tag{2}$$

It is helpful to first consider the case where p and q are constant, so we also introduce:

The strong LOT model hypothesis. *The strategy that the fielder uses to catch a fly ball is to follow a path that keeps both p and q constant.*

For either hypothesis, the line HI^* has slope p, so it follows that the line B^*F has slope $-1/p$; see Figure 2. Therefore, using the definition of slope, we have

$$-\frac{1}{p} = \frac{y_f - y_b}{x_f - x_b} \Rightarrow p = \frac{x_b - x_f}{y_f - y_b}, \tag{3}$$

and (3) is true at every point in time.

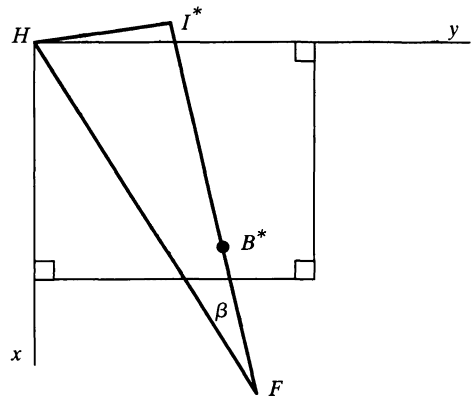


Figure 2

Next, we determine x_i in terms of B and p. The equation of the line HI^* is $y = px$, and the equation of the line B^*F is $y = y_b - (x - x_b)/p$. The point I^* is determined by the intersection of these two lines, and by setting these equations equal to each other and solving, we get

$$x_i = \frac{x_b + py_b}{p^2 + 1}. \tag{4}$$

Since F, B, and I are collinear, we have $(z_i - z_f)/(x_i - x_f) = (z_b - z_f)/(x_b - x_f)$, and since $z_f = 0$ and $z_i = qx_i$, solving this equation for x_f gives

$$x_f = x_i\left(\frac{z_b - qx_b}{z_b - qx_i}\right). \tag{5}$$

Combining (4) and (5) and simplifying yields

$$x_f = \frac{(z_b - qx_b)(x_b + py_b)}{z_b(p^2 + 1) - q(x_b + py_b)}. \tag{6}$$

Then, combining (3) and (6) and simplifying leads to

$$y_f = \frac{(pz_b - qy_b)(x_b + py_b)}{z_b(p^2 + 1) - q(x_b + py_b)}. \tag{7}$$

Finally, if we solve (6) for q, we get

$$q = \left(\frac{z_b}{x_b + py_b}\right)\left(\frac{x_f(p^2 + 1) - (x_b + py_b)}{x_f - x_b}\right). \tag{8}$$

What we now have, for every $t > 0$ and for every trajectory B, is a relationship between (x_f, y_f) and (p, q). If we know p and q, then (6) and (7) yields x_f and y_f, and if we know x_f and y_f, then (3) and (8) give us p and q. In the next two sections, we use these relationships to investigate the subtle elements of the LOT model.

3. THE STRONG LOT MODEL HYPOTHESIS: CONSEQUENCES AND PROBLEMS. For a given ball trajectory B, let T be the time when the ball is either caught or hits the ground, so that the ball is in flight for $0 < t < T$ and $z_b = 0$ when $t = T$. We assume that $x_b + py_b \nrightarrow 0$ as $t \to T$.

Theorem 3.1. *For a given ball trajectory B, and for every t_0 such that $0 < t_0 < T$, there exists a unique fielder's path, depending on the position of the fielder at time t_0, such that the fielder can use the strong LOT model for $t_0 \le t \le T$ on that path and catch the ball at time T.*

Proof of Theorem 3.1: At time t_0, the positions of the fielder and the ball are known. Therefore, using (3) and (8), we can determine p and q. Once these two constants are known, (6) and (7) specifies the unique path the fielder follows. At time T, we have $z_b = 0$, and therefore

$$x_f|_{t=T} = \left.\frac{(z_b - qx_b)(x_b + py_b)}{z_b(p^2+1) - q(x_b + py_b)}\right|_{t=T} = \left.\frac{-qx_b(x_b+py_b)}{-q(x_b+py_b)}\right|_{t=T} = x_b|_{t=T}. \quad (9)$$

Similarly, $y_f|_{t=T} = y_b|_{t=t}$. ■

Thus, the strong LOT model works, provided, of course, that the fielder can run fast enough to follow his predetermined path. Can a fielder track a ball starting at the moment the ball is launched by the batter? McBeath and his colleagues write, "...fielders do not and cannot arbitrarily select optical angles and rates of change... but rather they maintain the initial optical projection angle, Ψ, which is fully determined by the perspective launch angle of the ball relative to the fielder." [2] This suggests that formulas can be developed for p and q, and hence Ψ, from the information at $t = 0$, and that perhaps the strong LOT strategy can be utilized from the moment the ball is hit.

Lemma 3.1. *Under the assumption that p and q are constant near $t = 0$, the values of p and q are uniquely determined at $t = 0$ by the initial velocity of the ball and the initial position of the fielder.*

Proof of Lemma 3.1: Since (3) is true for all t, it is true for when the batter hits the baseball ($t = 0$), therefore we can conclude that

$$p|_{t=0} = \left.\frac{x_b - x_f}{y_f - y_b}\right|_{t=0} = -\left.\frac{x_f}{y_f}\right|_{t=0}. \quad (10)$$

To determine q, we use (8) and L'Hôpital's rule:

$$\begin{aligned} q|_{t=0} &= \lim_{t\to 0}\left(\frac{z_b}{x_b+py_b}\right)\left(\frac{x_f(p^2+1) - (x_b+py_b)}{x_f - x_b}\right) \\ &= \lim_{t\to 0}\left(\frac{z_b}{x_b+py_b}\right)(p^2+1) = \left.\frac{z_b'(p^2+1)}{x_b' + py_b'}\right|_{t=0}. \end{aligned} \quad (11)$$

Thus, (10) and (11) provide formulas for p and q as functions of the position of the fielder and the velocity vector of the ball at $t = 0$. ■

Lemma 3.1 clearly applies to the strong LOT model, but, as we see in the following theorem, we now have a problem with the strong LOT strategy at $t = 0$, because q does not depend on the initial position of the fielder.

Theorem 3.2. *For a given ball trajectory B, there exist points $(\bar{x}_f, \bar{y}_f, 0)$, called* ideal fielders' positions, *such that a fielder situated at that position when the ball is hit can*

use the strong LOT model for $0 \leq t \leq T$ to determine a unique path to catch the ball. Not all fielders' positions are ideal fielders' positions.

Proof of Theorem 3.2: Choose an arbitrary value of p. From Lemma 3.1, we can determine q as a function of B and p and define

$$\bar{x}_f = \lim_{t \to 0} \frac{(z_b - qx_b)(x_b - py_b)}{z_b(p^2 + 1) - q(x_b + py_b)} \tag{12}$$

and

$$\bar{y}_f = \lim_{t \to 0} \frac{(pz_b - qy_b)(x_b + py_b)}{z_b(p^2 + 1) - q(x_b + py_b)}. \tag{13}$$

A fielder situated at $(\bar{x}_f, \bar{y}_f, 0)$ at $t = 0$ can then use the unique path described by (6) and (7) in order to catch the ball at time T.

Now fix $\lambda \neq 1$ and consider a second fielder situated at $(\lambda \bar{x}_f, \lambda \bar{y}_f, 0)$ when $t = 0$. This fielder has the same p and q as the first fielder (due to Lemma 3.1), and therefore the same running path as the first fielder, which is impossible. The second fielder is not in an ideal fielder's position. ■

As a practical matter, since q depends on the velocity vector of the ball, a fielder would need some period of time after the ball was hit to recognize what Ψ is. This means an ideal fielder is ideal in another sense, because at $t = 0$, he can instantaneously discern the velocity of the baseball and begin his path to catch the ball. Theorem 3.1, which doesn't apply when $t = 0$, is a better description of how the strong LOT model operates, and it is simply incorrect to say that a fielder can use the LOT strategy from the crack of the bat. It makes more sense to say that fielders maintain a constant Ψ, which is determined a moment after the ball is launched by the batter.

4. ONE PATH OR MANY: USING MATHEMATICS TO CLARIFY IDEAS. In the case where we use the LOT model and not the strong LOT model, there is no longer a unique path that a fielder must follow in order to achieve a linear optical trajectory.

Theorem 4.1. *For a given ball trajectory B, and for every t_0 such that $0 < t_0 < T$, there exist an infinite number of fielders' paths, such that a fielder can use the LOT strategy for $t_0 \leq t \leq T$ on that path and catch the ball at time T.*

Proof of Theorem 4.1: This theorem is proved the same way as Theorem 3.1, except now, as t increases, both p and q are allowed to vary, as long as, by Lemma 2.1, $q^2/(1 + p^2)$ remains constant. This gives an infinite number of solutions. ■

With the assumption of Lemma 3.1, we can also define ideal fielders' positions for the LOT model. The proof is the same as Theorem 4.2, except that now the path determined by (12) and (13) is not unique.

We can use our results to examine the validity of, and explain, several statements from [1] and [2]. The benefit of the mathematical analysis is that we can recognize how these statements follow from the LOT model and gain a greater

appreciation for the model. Here are the statements:

> A. ... the [LOT model] strategy in itself does not specify a unique solution. [2]
> B. ... angle of bearing appears to be used as an additional constraint to help determine the particular LOT chosen. [2]
> C. "[The LOT model] keeps the image of the ball continuously ascending in a straight line throughout the trajectory." [1]
> D. One interesting aspect that has emerged from research on this problem is that for identical launches, fielders will select different running paths, particularly near the beginning and end of the task. A good model of outfielder behavior should allow for this variability, as the LOT strategy does. Near the beginning of the trajectory [of the ball], we expect more variability because outfielder location has less influence on the optical trajectory. Near the end we expect more variability because corrective action will commence as other depth cues become available. [2]

Statement A is a result of Theorem 4.1. The angle of bearing in statement B is a function of p and q, and since q is a function of p by Lemma 2.1, it makes sense to call p the *fielder's bearing function*. As the fielder tracks the ball, he unconsciously chooses a function p as a part of his LOT strategy. (The strong LOT model keeps the angle of bearing constant.) Since it is reasonable to assume that a player's bearings will not change in the first moment after the ball is hit, we are justified in keeping p constant near $t = 0$ in Lemma 3.1.

Statement C seems at odds with the idea that p and q can be allowed to vary, which suggests that the LOT hypothesis does not allow the trajectory of I to be non-linear. A fielder's bearings *can* change, and this corresponds to a rotation about home plate of the right pyramid in Figure 1. For example, if the bearing function changes from a value of p_1 to a value of p_2, then the angle of rotation is $\theta = \arctan p_2 - \arctan p_1$. From the rotation of axes formulas, we get

$$\cos\theta = \frac{1 + p_1 p_2}{\sqrt{(1+p_1^2)}\sqrt{(1+p_2^2)}} \quad \text{and} \quad \sin\theta = \frac{p_2 - p_1}{\sqrt{(1+p_1^2)}\sqrt{(1+p_2^2)}}. \tag{15}$$

Suppose we have a situation where first we have $I_1 = (x_i, p_1 x_i, q_1 x_i)$ and then a change of bearings leading to $I_2 = (x_i, p_2 x_i, q_2 x_i)$. Applying the change of variables formula to I_2, we end up with the rotated point

$$I_2' = \sqrt{\frac{1+p_2^2}{1+p_1^2}}\,(x_i, p_1 x_i, q_1 x_i). \tag{16}$$

From the perspective of the fielder, the ball appears to remain on the vector $\langle 1, p_1, q_1 \rangle$, "continuously ascending on a straight line."

Statement D is a consequence of the proof of Theorem 4.1, which explains how there can be many paths that keep Ψ constant. Also, from (3), we get

$$\frac{\partial p}{\partial x_f} = -\frac{1}{y_f - y_b} \quad \text{and} \quad \frac{\partial p}{\partial y_f} = -\frac{x_b - x_f}{(y_f - y_b)^2}. \tag{17}$$

These derivatives are small when t is near 0, as are the derivatives for q. As a consequence, when (3) and (8) are used to determine p and q "near the beginning of the trajectory," fielders near each other have quite similar optical trajectory projection angles Ψ, hence "outfielder location has less influence on optical trajectory." Therefore, different running paths can keep p and q nearly constant, as the LOT model predicts.

The mathematics presented here shows that the LOT model is reasonable and, interestingly, qualitative observations made by the researchers can be supported quantitatively by the analysis. This analysis, though, *does not prove* that the LOT model is correct. The LOT model was developed by perceptual psychologists using statistical methods and it cannot be proved as we prove a theorem in mathematics. It also goes without saying that outfielders do not and cannot follow the LOT strategy *exactly*, or even that outfielders follow the strategy well. In **[1]**, McBeath et al. reported that one fielder appeared to use a linear optical trajectory for a while, faltered, then continued on a new optical trajectory with a different Ψ!

There are other, competing models, such as the optical acceleration cancellation (OAC) model, that have their defenders. According to the OAC model, a fielder acts to keep $d(\tan \alpha)/dt$ constant, not Ψ. Another view comes from Robert K. Adair, a physicist, who argues that "a fielder runs laterally so that the ball goes straight up and down from his or her view." **[3]** The OAC model or Adair's may be correct, although McBeath and his colleagues rebutted both theories with this statement, "Both maintenance of lateral alignment and monitoring of up and down ball motion require information that is not perceptually available from the fielder's vantage." **[2]**

An example of the failure of the LOT strategy was recently presented by James L. Dannemiller, Timothy G. Babler, and Brian L. Babler. They write that it is possible for a fielder to use the LOT strategy and arrive "away from the ball's landing site at the instant the ball hits the ground." **[5]** This *is* possible if the fielder chooses a path such that $x_b + py_b \to 0$ as $t \to T$. In this case, Theorems 3.1 and 4.1 are invalid and $B \neq F$ when $t = T$. However, in this instance, H, B, and F would become collinear at the moment the ball hits the ground.

5. A MATHEMATICIAN CATCHES A BASEBALL. We now go to the ball park, and a mathematician on the visiting team is standing in right field, waiting to catch some fly balls for his team. It is the bottom of the ninth inning, and the visitors are ahead 4–3. The ball is hit! Let's suppose that B^*, the projection of the ball on the field, moves in a straight line and that the path of the ball is a parabolic arc. These are reasonable assumptions (unless we are playing in Wrigley Field, where it is rather windy at times) and an example of a ball's trajectory (in feet) would be

$$x_b(t) = 75t, \qquad y_b(t) = 10t, \qquad z_b(t) = -64t^2 + 256t. \tag{18}$$

This is a ball hit to deep right field that will be in the air for 4 seconds and will land approximately 303 feet from home plate unless our intrepid mathematician gets there in time.

Suppose that our right fielder is positioned in right-center field at the crack of the bat at the position $(x_f, y_f, z_f) = (270, 70, 0)$, that it takes him 0.3 seconds to get his bearings, and that he plans to use the strong LOT model to catch the ball. Utilizing Theorem 3.1, he determines that

$$p \approx -3.694 \qquad \text{and} \qquad q \approx 99.121. \tag{19}$$

Therefore, $\Psi \approx 87.8°$; this is practically vertical—perhaps Adair has a point! Now that p and q are known, a unique path can be determined for our all-star to catch this fly ball. That path is indicated in Figure 3.

On the very next play, inexplicably another ball is hit to right field, with the same trajectory and the same response time for our athlete. This time, though, he decides to use the regular LOT strategy, computing his initial p and q as above, and then using $p = -0.0827t - 3.6692$ as his bearing function. This path, which

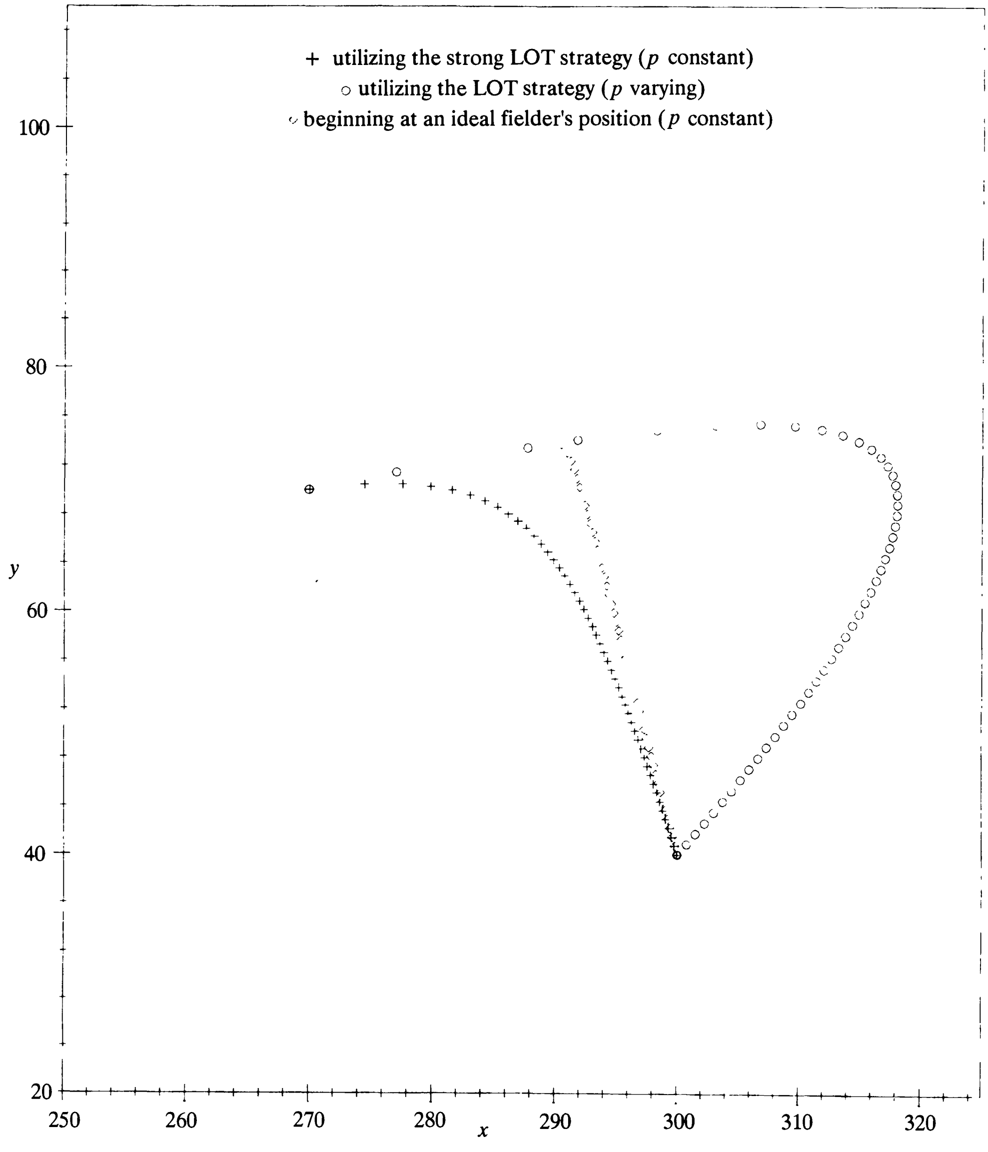


Figure 3

sends the fielder to the edge of the playing field and back, is also indicated in Figure 3.

Now that there are two outs, our mathematician wonders if the next batter will also hit a fly ball satisfying (18). He decides to get into an ideal fielder's position. Although the most obvious one is (300, 40, 0)—the point where the ball has landed the first two times—he decides to find an ideal position that corresponds to his current position. He computes $p \approx -3.857$, using (10), and $q \approx 111.579$, using (8). Then, making use of equations (12) and (13), and a little calculus, he determines

$$\bar{x}_f \approx 291 \quad \text{and} \quad \bar{y}_f \approx 75. \tag{20}$$

Figure 4 shows, for a ball following trajectory (18), several ideal fielder's positions, depending on the choice of p. Again in Figure 3, we see out hero,

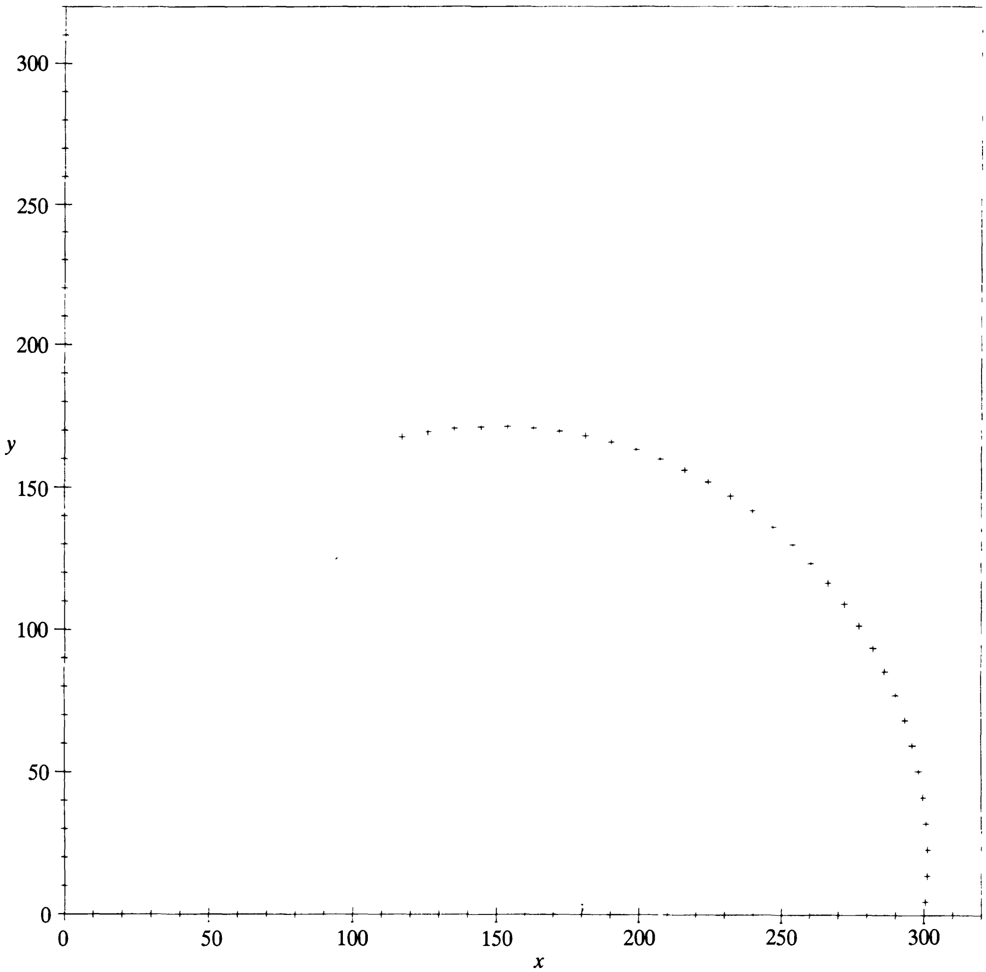


Figure 4

starting at (20), pursuing the ball using the strong LOT strategy and catching it after 4 seconds. That's three outs—game over!

REFERENCES

[1] M. McBeath, D. Shaffer, and M. Kaiser, How Baseball Outfielders Determine Where to Run to Catch Fly Balls, *Science* **268** (28 April 1995), 569–573.

[2] M. McBeath, D. Shaffer, and M. Kaiser, Letters, *Science* **268** (23 June 1995), 1683–1685.

[3] R. Adair, Letters, *Sciences* **268** (23 June 1995), 1681–1682.

[4] P. Hilts, New Theory Offered on How Outfielders Snag Their Prey, *The New York Times*, 28 April 1995, B9.

[5] J. Dannemiller, T. Babler, and B. Babler, On Catching Fly Balls, *Science* **273** (12 July 1996), 256–257.

Department of Mathematics & Statistics
Grand Valley State University
Allendale, MI 49401-9403
aboufade@gvsu.edu

Hilbert's 17th Problem and the Champagne Problem

Victoria Powers

Dedicated to Alex Rosenberg on the occasion of his 70th birthday

INTRODUCTION. About 15 years ago, E. Becker gave a talk in which he proved that

$$B(t) := \frac{1+t^2}{2+t^2} \in \mathbb{Q}(t)$$

is a sum of $2n$-th powers of elements in $\mathbb{Q}(t)$ for all n. To prove this surprising fact, Becker used his newly developed theory of higher level orderings on fields; the proof was not constructive. He then proposed the following problem: Find an explicit formula giving a representation of $B(t)$ as a sum of $2n$-th powers (in $\mathbb{Q}(t)$) for all n. Becker promised a bottle of champagne to the first person to solve this; as a result, the problem became known as the Champagne Problem. The problem still remains unsolved in the form stated by Becker; however, recent work of B. Reznick gives an explicit formula for $B(t)$ as a sum of $2n$-th powers of elements in $\mathbb{R}(t)$.

The theory of higher level orderings on fields, and hence the Champagne Problem, has its genesis in Hilbert's 17th Problem and E. Artin's solution to it. In this paper, we trace the history of these roots of the Champagne Problem and briefly describe Reznick's solution.

FROM HILBERT'S 17TH PROBLEM TO ORDERED FIELDS. The Champagne Problem is part of a class of problems concerned with representations of positive semi-definite rational functions as sums of squares of rational functions or, more generally, sums of even powers. A rational function $f \in \mathbb{R}(X) := \mathbb{R}(x_1, \dots, x_k)$ is *positive semi-definite* (psd) if $f \geq 0$ at every point in $\mathbb{R}^k$ for which it is defined. D. Hilbert, in a paper published in 1888 **[22]**, showed that there exist psd polynomials that are not sums of squares of polynomials. In 1900, at the International Congress of Mathematics in Paris, Hilbert gave a lecture in which he proposed 23 open problems. Most of these have since been solved, and the solutions have led to fundamental discoveries in mathematics.

Hilbert's work on sums of squares of polynomials was the impetus for the 17th problem: "A rational integral function or form in any number of variables with real coefficients such that it becomes negative for no real values of these variables, is said to be *definite*. But since, as I have shown, not every definite form can be compounded by addition from squares of forms, the question arises—which I have

answered affirmatively for ternary forms ([**23**])—whether every definite form may not be expressed as a quotient of sums of squares of forms" [**24**]. In other words, given a psd polynomial $f \in \mathbb{R}[x_1, \ldots, x_k]$, can f be written as a sum of squares of elements in $\mathbb{R}(X)$?

In addition to Hilbert's work on sums of squares, the 17th problem arises from questions in elementary geometry involving ruler and compass constructions. This relationship is described in Hilbert's book on foundations of geometry [**26**].

In 1927, Artin solved the 17th Problem in the affirmative [**1**], using the theory of ordered fields. Artin realized that it is important to study sums of squares in arbitrary fields. For a field F, let $\Sigma F^2 = \{y_1^2 + \cdots + y_r^2 | r \in \mathbb{N}, y_i \in F\}$. If F has characteristic 2, then $\Sigma F^2 = F^2$, and if char $F \neq 2$ and $-1 \in \Sigma F^2$, it follows from the formula $x = ((x+1)/2)^2 - ((x-1)/2)^2$ that $\Sigma F^2 = F$. This leaves the case $-1 \notin \Sigma F^2$; such fields are called *formally real*. Notice that if F is formally real, then F must have characteristic 0 since char $F = n > 0$ would imply $-1 = (n-1) \cdot 1 \in \Sigma F^2$. Artin, along with O. Schreier, showed that formally real fields are precisely those that have an order.

The definition of an ordered field goes back to Hilbert, who did not develop the theory. It was Artin and Schreier who laid out the fundamentals of the theory in two papers published in 1927 [**2**], [**3**]. We briefly describe their major results.

An *order* on a field F is given by a subset $P \subseteq F$, sometimes called the *positive cone* of the order, which satisfies: $P \cdot P \subseteq P$, $P + P \subseteq P$, $P \cap -P = \{0\}$, and $P \cup -P = F$. We will write "P is an order" to mean P is the positive cone of an order. If P is a given order on F, we can define binary relations $\leq$ and $<$ on F by $x \leq y$ if and only if $y - x \in P$, and $x < y$ if and only if $y - x \in P$ and $y \neq x$. One can check that $<$ is then a total order on F and the usual rules for inequalities hold for $<$ and $\leq$. Note that we can recover P from $\leq$ via $P = \{x \in F | 0 \leq x\}$.

The easiest examples of orders are the obvious orders on $\mathbb{Q}$ and $\mathbb{R}$, and it is not too hard to show that these are the only orders on these fields. Suppose F is a subfield of K, which has an order P; it is easy to check that $F \cap P$ is an order on F. However, not all orders on F arise in this fashion. To see this, consider $F = \mathbb{Q}[\sqrt{2}]$; in addition to the order arising from the order on $\mathbb{R}$, one can check that $\{a + b\sqrt{2} | 0 \leq a - b\sqrt{2} \text{ in } \mathbb{R}\}$ is also an order.

For a subset $S \subseteq F$, we write $\dot{S}$ to denote $S \setminus \{0\}$. We can show that $\Sigma \dot{F}^2$ is a subgroup of $\dot{F}$ using the fact that if $y = \Sigma y_i^2 \in \Sigma \dot{F}^2$, then $1/y = \Sigma (y_i/y)^2 \in \Sigma \dot{F}^2$. Given any order P on F and $x \in F$, since $x \in P$ or $-x \in P$, it follows that $x^2 = (\pm x)(\pm x) \in P$. Hence, by additive closure, $\Sigma F^2 \subseteq P$. Then $\dot{P}$ is a subgroup of $\dot{F}$, since $x \in \dot{P}$ implies $1/x = x \cdot 1/x^2 \in \dot{P} \cdot \dot{P} \subseteq \dot{P}$. Note that $\dot{P}$ is a subgroup of index 2 in $\dot{F}$ that is additively closed. It is easy to see that, conversely, for any subgroup Q of index 2 in $\dot{F}$ that is additively closed, $Q \cup \{0\}$ is an order on F.

For any order P, we have just seen that $\Sigma F^2 \subseteq P$. Thus $1 \in P$ and hence, if F has an order, F must be formally real. Artin and Schreier proved the converse:

Artin-Schreier Theorem. *If a field F is formally real, then F admits an order.*

The key idea needed for the proof of this theorem is that of a real closed field: F is *real closed* if F is formally real and no proper algebraic extension of F is formally real. Note that by Zorn's Lemma, any formally real field admits a maximal algebraic extension that is formally real, hence every formally real field is con-

tained in a real closed field. Artin and Schreier proved the following characterization of real closed fields:

Theorem. *For a field F the following are equivalent*:

(i) *F is real closed*,

(ii) $-1 \notin F^2$ *and* $F[\sqrt{-1}]$ *is algebraically closed*,

(iii) *F is formally real, $\dot{F}^2$ has index 2 in $\dot{F}$ and is additively closed, and any polynomial of odd degree in $F[x]$ has a root in F.*

Using this theorem, the proof of the Artin-Schreier theorem is easy: Given a formally real field F, then as stated above, F is contained in a real closed field R. From (iii) of the theorem, the set R^2 is an order on R, hence $R^2 \cap F$ is an order on F.

The Fundamental Theorem of Algebra can also be recovered from the theorem: Since $\mathbb{R}$ satisfies (iii), $\mathbb{R}$ is real closed and then, by (ii), $\mathbb{C}$ is algebraically closed.

We note in passing the amazing "meta-theorem" of A. Tarski, now known as Tarski's Principle. Before stating it, we define a *formula of the language of ordered fields* as any formula expressible using field operations, inequalities, and the logical symbols $\vee$ (disjunction), $\wedge$ (conjunction), negation, and quantifiers. Then Tarski's Principle says that any formula in the language of ordered fields that holds over $\mathbb{R}$, holds over every real closed field **[48]**.

Artin proved the following theorem relating orders in F to sums of squares:

Theorem. *If F is formally real, then $\Sigma F^2 = \bigcap P$, where the intersection is over all orders P in F.*

Let's verify the theorem for $F = \mathbb{Q}[\sqrt{2}]$. First we claim that the only orders on F are the two mentioned above: $P_1 := \{a + b\sqrt{2} | a + b\sqrt{2} \geq 0 \text{ in } \mathbb{R}\}$ and $P_2 := \{a + b\sqrt{2} | a - b\sqrt{2} \geq 0 \text{ in } \mathbb{R}\}$. If P is an order, then $\mathbb{Q}^+ \subseteq P$ and $\sqrt{2} \in \pm P$. If $\sqrt{2} \in P$, then $P_1 \subseteq P$, which implies that $\dot{P} = \dot{P}_1$ since both have index 2 in $\dot{F}$. Hence $P = P_1$. If $-\sqrt{2} \in P$, then a similar argument shows $\dot{P} = \dot{P}_2$ and thus $P = P_2$. Now $P_1 \cap P_2 = \{a + b\sqrt{2} | a \geq 0 \text{ and } a^2 \geq 2b^2\}$. We can show directly that this is precisely ΣF^2. For any $c, d \in \mathbb{Q}$, we have $(c + d\sqrt{2})^2 = (c^2 + 2d^2) + 2cd\sqrt{2} \in P_1 \cap P_2$, hence $\Sigma F^2 \subseteq P_1 \cap P_2$. Now suppose $a + b\sqrt{2} \in P_1 \cap P_2$. If $b = 0$, then $a \geq 0$ in $\mathbb{Q}$, hence $a \in \Sigma \mathbb{Q}^2 \subseteq \Sigma F^2$. If $b < 0$ and we can write $a - b\sqrt{2}$ as a sum of squares, then taking "conjugates" we can do it for $a + b\sqrt{2}$, so we may assume $b > 0$. Consider the square $(x + (b/2x)\sqrt{2})^2 = x^2 + b^2/2x^2 + b\sqrt{2}$. As a function of x^2, $x^2 + b^2/2x^2$ attains its minimum at $x^2 = b/\sqrt{2}$ and has value $b\sqrt{2} \leq a$. Hence we can find a rational q so that $q^2 + b^2/2q^2 \leq a$. Then $a + b\sqrt{2} = (q + (b/2q)\sqrt{2})^2 + (a - (q^2 + b^2/2q^2)) \in \Sigma F^2$. Thus we have verified the theorem for the case of $\mathbb{Q}[\sqrt{2}]$.

It remained for Artin to show that if $f \in \mathbb{R}(X)$ is psd then f is in every order on $\mathbb{R}(X)$, hence is a sum of squares in $\mathbb{R}(X)$. In fact, Artin proved more than this. He showed that given any subfield $F \subseteq \mathbb{R}$ that has only one order and any $f \in F(X) := F(x_1, \ldots, x_k)$ such that $f \geq 0$ (in the unique ordering on F) at every point at which it is defined, then f is in every order of $F(X)$. To show this, Artin proved a series of "specialization lemmas" by using Sturm's Theorem, which is an algorithm for counting exactly the number of real roots of a polynomial, see **[46]**. For a modern exposition of Sturm's Theorem and Artin's proof, see **[35**, Chap. XI §2].

FROM ORDERS TO HIGHER LEVEL ORDERINGS. In the late 1970's, Becker discovered a generalization of the notion of an order on a field, which led to a natural, far-reaching extension of the Artin-Schreier theory. As stated above, if $P \subseteq F$ is an order, then $\dot{P}$ is a subgroup of $\dot{F}$ of index 2 that is additively closed. Furthermore, this description is equivalent to the previous definition of order. An *ordering of higher level* on a field F is a subset $P \subseteq F$ (containing 0) such that $\dot{P}$ is an additively closed subgroup of $\dot{F}$ for which $\dot{F}/\dot{P}$ is finite cyclic. Since $\dot{P}$ is additively closed, $-1 \notin \dot{P}$ and hence $\dot{P}$ must have even index in $\dot{F}$. The *level* of P is $\frac{1}{2}[\dot{F} : \dot{P}]$. Note that ordinary orders are simply orderings of level 1; in other words, instead of requiring that $\dot{P}$ have index 2 in $\dot{F}$, Becker required it to have index $2n$.

For an example of a field with orderings of all levels, consider $\mathbb{R}((x))$, the field of formal power series in x over $\mathbb{R}$. Elements are formal series $\sum_{i=m}^{\infty} \alpha_i x^i$, where $m \in \mathbb{Z}$ and $\alpha_i \in \mathbb{R}$. Fix $n \in \mathbb{N}$ and let z be a primitive $2n$-th root of 1. It is easy to check that $P_z := \{\sum_{i=m}^{\infty} \alpha_i x^i | \alpha_m z^m = 1\} \cup \{0\}$ is an ordering of level n.

Becker **[7]** obtained generalizations of the Artin-Schreier and Artin theorems:

Higher Level Artin-Schreier Theorem. *The following are equivalent for a field F and any $n \in \mathbb{N}$:*

(i) *F has an ordering of level n,*

(ii) $-1 \notin \Sigma F^{2n}$,

(iii) *F is formally real.*

Furthermore, if F is formally real, then for all n, $\Sigma F^{2n} = \bigcap P$, where the intersection is over all orderings of level dividing n.

It should be noted that before Becker proved this theorem, J. Joly **[27]** proved the equivalence of (ii) and (iii), without making use of the notion of a higher level ordering. In fact, Joly proved this for any commutative ring.

When working with orders and orderings, it is impossible to avoid valuation theory, since this is one of the main tools for studying formally real fields. The intimate connections between valuation theory and the theory of orders were first seen in the work of R. Baer **[5]**, **[6]** and W. Krull **[30]**, soon after the theory of ordered fields was developed.

A subring $V \subseteq F$ is a *valuation ring* if V contains x or x^{-1} for every nonzero $x \in F$. In this case, the set $\mathscr{M} = \{x \in \dot{V} | x^{-1} \notin V\} \cup \{0\}$ forms the unique maximal ideal in V and the field $V/\mathscr{M}$ is called the *residue field* of V. We say V is a *real valuation ring* if the residue field is formally real. One can show that F has a real valuation ring if and only if F is itself formally real.

Valuation theory arises naturally in ordered fields in the following way: Given an ordering P on F (of some level), let $A(P) = \{x \in F | q \pm x \in P$ for some $q \in \mathbb{Q}^+\}$, and let $I(P) = \{x \in F | q \pm x \in P$ for all $q \in \mathbb{Q}^+\}$. Then we have

Theorem. *$A(P)$ is a valuation ring in F with maximal ideal $I(P)$, and the set $\bar{P} := \{x + I(P) | x \in A(P) \cap P\}$ is an order in the residue field $A(P)/I(P)$.*

The proof of this theorem for orders is a straightforward calculation; however, the proof for higher level orderings is much harder. Becker makes use of the Kadison-Dubois representation theorem from functional analysis. If the ordering has level a power of 2, there is a direct proof due to A. Wadsworth (unpublished). The theorem is the key result Becker needed for the proof of the Higher Level Artin-Schreier Theorem. Given that result, the proof of the first part now proceeds

as follows: If F has an ordering of some level, then F has a real valuation ring, so F is formally real. Later work by Becker, J. Harman, and A. Rosenberg **[12]** showed that the orderings of F of all levels can be described using only the (level 1) orders in F and the real valuation rings in F.

Given a formally real field F, the *real holomorphy ring* of $F, H := H(F)$, is the intersection of all real valuation rings in F. Now let $\mathbb{E} := \mathbb{E}(F)$ denote the units in $H(F)$, and set $\mathbb{E}^+ := \mathbb{E} \cap \Sigma F^2$. Although these definitions seem at first to have little to do with the higher level theory, in **[9]** and **[10]** Becker shows that there is an intimate connection between the structure of $\mathbb{E}^+$ and ΣF^{2n}:

Theorem. *Let F be formally real and let H and $\mathbb{E}^+$ be as above. Then*

(i) $\mathbb{E}^+ = \{r(s+q)/(t+q) | r, s, t \in \mathbb{Q}^+, q \in \Sigma F^2\}$,

(ii) $\mathbb{E}^+ \subseteq \bigcap_{n \in \mathbb{N}} \Sigma F^{2n}$, *and*

(iii) *For all $n \in \mathbb{N}$, $\Sigma F^{2n} = \mathbb{E}^+ \cdot (\Sigma F^2)^n$.*

It follows from (i) and (ii) that

$$B(t) = \frac{1+t^2}{2+t^2} \in \sum \mathbb{Q}(t)^{2n} \text{ for all } n$$

and hence we arrive at the Champagne Problem: Find an explicit formula expressing $B(t)$ as a sum of $2n$-th powers (in $\mathbb{Q}(t)$) for all n.

The preceding theorem allows one to construct many examples of sums of $2n$-th powers. Further, for certain fields including $\mathbb{R}(X)$, Becker shows that $\mathbb{E}^+ = \bigcap_{n \in \mathbb{N}} \Sigma F^{2n}$.

Notice that it follows from (iii) of the theorem that for any n,

$$\left(\sum F^2\right)^n \subseteq \sum F^{2n},$$

a highly non-obvious fact! For example, over $\mathbb{R}$ we have

$$(*) \qquad (x^2+y^2)^3 = \frac{4}{5}\left(x^6 + \left(\frac{x+y}{\sqrt{2}}\right)^6 + y^6 + \left(\frac{-x+y}{\sqrt{2}}\right)^6\right),$$

which can be checked by hand.

Using the higher level theory, Becker extends this to show that given $n, m \in \mathbb{N}$, there exist identities

$$(**) \qquad \left(x_1^{2n} + \cdots + x_k^{2n}\right)^m = f_1^{2nm} + \cdots + f_r^{2nm},$$

where $f_i \in \mathbb{Q}(x_1, \ldots, x_k)$. For details, see **[8]**. It should be noted that Becker showed only the existence of the identities; they are not given in any explicit way. Hilbert proved the existence of identities $(**)$ in the case $n = 1$; moreover, in this case he showed that the f_i's can be chosen to be polynomials, see **[25]**. For this reason, when $n = 1$ and the f_i's are polynomials, the identities $(**)$ are usually called Hilbert Identities. We shall see that the Hilbert Identities, or, more precisely, an explicit version of them over $\mathbb{R}$, play a key role in Reznick's solution to the Champagne Problem.

THE CHAMPAGNE PROBLEM SOLVED. Artin's solution to Hilbert's 17th problem is not constructive, nor are the proofs of Becker's theorems on sums of even powers in formally real fields. Thus a natural question that arises from the 17th Problem is to what extent can we find an explicit representation of a psd f as a sum of squares of rational functions or, more generally, as a sum of $2n$-th powers

for any n. The Champagne Problem is thus a specific example of this type of question. We discuss some computational aspects of the 17th Problem, and conclude with a sketch of Reznick's solution (over $\mathbb{R}$) to the Champagne Problem.

A year after Artin's solution to the 17th problem, G. Pólya [**38**] (see also [**21**, pp. 57–59]) found an explicit solution in a special case. He showed that if $f \in \mathbb{R}(X)$ is positive definite, i.e., $f(a) > 0$ for all $a \in \mathbb{R}^k$, and even (as a function), then for large enough r, $f \cdot (\Sigma x_i^2)^r$ is a sum of squares of monomials; in particular, f is a sum of squares of rational functions with common denominator $(\Sigma x_i^2)^r$. Recent work of J. de Loera and F. Santos [**17**] gives algorithms for finding a representation and bounds for r. In 1940, W. Habicht [**20**], using Pólya's result, showed directly that any positive definite polynomial f can be written as a sum of squares of rational functions, and that if f has only rational coefficients then so do the monomials. Recently, Reznick [**42**] has extended this to show that any positive definite f can be written as a sum of squares of rational functions with common denominator $(\Sigma x_i^2)^r$. In other words, Reznick has extended Pólya's result to the more general setting of Habicht's result. The computations used for this result enabled Reznick to obtain a representation for $B(t)$ as a sum of $2n$-th powers in $\mathbb{R}(t)$.

As mentioned in the previous section, a key idea needed is the Hilbert Identities. Let us state them more precisely: Given k and s, let $N = \binom{k+2s-1}{k-1}$. Then there exist $\alpha_{ij} \in \mathbb{Q}$ and positive $\lambda_i \in \mathbb{Q}$ such that

$$\left(x_1^2 + \cdots + x_k^2\right)^s = \sum_{i=1}^{N} \lambda_i \left(\alpha_{i1} x_1 + \cdots + \alpha_{ik} x_k\right)^{2s}.$$

There are no known explicit formulas except in the cases $s = 1, 2, 3$ (see [**41**, §8, §9]). However, if we allow formulas over $\mathbb{R}$ instead of $\mathbb{Q}$, we can obtain the following: If s and v are positive integers and $v \geq s + 1$, then

$$(\dagger) \qquad \left(x^2 + y^2\right)^s = \frac{2^{2s}}{v\binom{2s}{s}} \sum_{j=0}^{v-1} \left(\cos\left(\frac{j\pi}{v}\right)x + \sin\left(\frac{j\pi}{v}\right)y\right)^{2s}.$$

For a proof of this, see [**41**, 9.5]. Notice that the identity $(*)$ is $(\dagger)$ with $s = 3$ and $v = 4$.

Consider the following obvious equality:

$$B(t) = \frac{1 + t^2}{2 + t^2} = \frac{(1 + t^2)(2 + t^2)^{2n-1}}{(2 + t^2)^{2n}}$$

Then if we can write $(1 + t^2)(2 + t^2)^{n-1}$ and $(2 + t^2)^n$ as sums of $2n$-th powers of polynomials, their product is a sum of $2n$-th powers, each of which can be divided by $(2 + t^2)^{2n}$ to give $B(t)$ as a sum of $2n$-th powers of rational functions. Taking $s = n$, $v = n + 2$, $x = \sqrt{2}$, and $y = t$ in $(\dagger)$, after several pages of calculations Reznick obtains the remarkable formula

$$B(t) = \frac{2^{4n-2}}{n(n+2)^2\binom{2n}{n}^2} \sum_{i=0}^{n+1} \sum_{j=0}^{n+1} \lambda_j \left(\frac{L_i(\sqrt{2}, t) L_j(\sqrt{2}, t)}{2 + t^2}\right)^{2n},$$

where $\lambda_j = 3n - (n + 1)\cos(2j\pi/(n + 2))$ and $L_j(x, y) = (\cos(2j\pi/(n + 2))x + (\sin(2j\pi/(n + 2))y$. Thus we have an explicit formula for writing $B(t)$ as a sum of

$2n$-th powers in $\mathbb{R}(t)$, in other words, a solution to the Champagne Problem over the reals!

Unfortunately, there are no known explicit formulas for the Hilbert Identities in general; hence this method cannot produce a solution over $\mathbb{Q}$. However, Reznick's formula was close enough to a solution to the original problem that when he gave a talk on this work at the AMS/MAA joint winter meeting in Cincinnati in 1994, he was presented (by proxy) with a bottle of champagne from Becker.

ACKNOWLEDGEMENTS AND SUGGESTIONS FOR FURTHER READING. Thanks to Bruce Reznick for many helpful comments and suggestions. His recent paper **[43]** contains an exposition of Hilbert's work on sums of squares of polynomials that led to the 17th Problem, detailed information on sums of squares in general, and much more. Our sketch of the derivation of the formula for $B(t)$ is paraphrased from **[43]**. We have also been influenced by other papers in this area, particularly Lam's excellent expository article on ordered fields **[32]**.

The text of Hilbert's 1900 lecture at the ICM can be found in **[15]**, along with descriptions of the mathematical developments arising from the 23 problems he proposed. For an up-to-date account of the status of the Hilbert problems, see the recent article by J.-M. Kantor **[28]**. Much has been written on the 17th Problem and mathematical developments arising from it. Here we mention a few articles that are well worth reading. A. Pfister **[37]** and P. Ribenboim **[44]** wrote surveys of the 17th Problem in the 70's. A more recent survey was written by D. Gondard **[19]**. See also C. Scheiderer's survey **[45]**, where connections with real algebra and applications to geometry are discussed. A recent article of C. Delzell **[18]** describes the history of the 17th Problem and its relationship to questions in logic. For a nice discussion of the relationship of the 17th problem to questions in elementary geometry, see the article of D. Auckly and J. Cleveland on origami and paper folding **[4]**.

For more on the theory of ordered fields, see **[32]**. The generalization of the notion of an order to commutative rings led to the development of real algebra and real algebraic geometry, see **[11]**, **[14]**, and **[34]**. Much of the Artin-Schreier theory has been generalized to semi-local rings by M. Knebusch, Rosenberg, and R. Ware **[29]**. The notion of an order and some of the theory was extended to division rings by T. Szele **[47]**. Much of the higher level theory for fields has also been extended, to commutative rings **[36]**, **[40]**, to division rings **[16]**, and even to general non-commutative rings **[39]**. The theory of ordered fields and the related valuation theory is intimately connected with the algebraic theory of quadratic forms; see Lam's books **[31]** and **[33]**. Work of Becker and Rosenberg shows that these connections also exist in the higher level theory, see **[13]**.

REFERENCES

[1] E. Artin, Über der zerlegung definiter funktionen in quadrate, *Hamb. Abh.* **5** (1927), 273–288.

[2] E. Artin and O. Schreier, Algebraische konstruktion reeller Körper, *Hamb. Abh.* **5** (1927), 85–99.

[3] E. Artin and O. Schreier, Eine Kennzeichnung der reell abgeschlossenen Körper, *Hamb. Abh.* **5** (1927), 225–231.

[4] D. Auckly and J. Cleveland, Totally real origami and impossible paper folding, *Amer. Math. Monthly* **102** (1995), 215–226.

[5] R. Baer, Über nicht-archimedisch geordnete Körper, *Sitz. Ber. der Heidelberger Akd.* **8** (1927), 3–13.

[6] R. Baer, Zur topologie der gruppen, *J. für die reine und angew. Math.* **160** (1929), 208–226.

[7] E. Becker, *Hereditarily-Pythagorean fields and orderings of higher level*, Lecture Notes, Instituto de Matematica Pura e Aplicada, Vol. 29, Rio de Janeiro, 1978.

[8] E. Becker, Summen n-ter potenzen in Körper, *J. für die reine und angew. Math.* **307 / 8** (1979), 8–30.
[9] E. Becker, *Valuations and real places in the theory of formally real fields*, Géométrie algébrique réelle et formes quadratiques, Proceedings, Rennes, 1981, Lecture Notes in Math. No. 959, Springer-Verlag, Berlin, Heidelberg, New York, 1982, pp. 1–40.
[10] E. Becker, *The real holomorphy ring and sums of 2n-th powers*, loc. cit. **[9]**, pp. 139–181.
[11] E. Becker, On the real spectrum of a ring and its application to semialgebraic geometry, *Bull. Amer. Math. Soc.* **15** (1986), 19–60.
[12] E. Becker, J. Harman, and A. Rosenberg, Signatures of fields and extension theory, *J. für die reine und angew. Math.* **330** (1982), 53–75.
[13] E. Becker and A. Rosenberg, Reduced forms and reduced Witt rings of higher level, *J. Algebra* **92** (1985), 477–503.
[14] J. Bochnak, M. Coste, and M.-F. Roy, *Géométrie Algébrique Réelle*, Springer-Verlag, Berlin, Heidelberg, New York, 1987.
[15] F. Browder, editor, *Mathematical Developments Arising from Hilbert Problems*, Proc. Symp. Pure Math., Vol. 28, Amer. Math. Soc., Providence, RI, 1976.
[16] T. Craven, Witt rings and orderings of skew fields, *J. Algebra* **77** (1982), 74–96.
[17] J. A. de Loera and F. Santos, An effective version of Pólya's theorem on positive definite forms, *J. Pure and Applied Algebra* **108** (1996), 231–240.
[18] C. Delzell, *Kreisel's unwinding of Artin's proof*, About and Around Georg Kreisel (P. Odifreddi, ed.), to appear, A. K. Peters, Wellesley, MA.
[19] D. Gondard-Cozette, *Le 17éme probléme de Hilbert et ses développements récents*, Seminaire sur les structures algébriques ordonnées: Sélection d'exposés 1984–1987, Vol. II, Publ. Math. Univ. Paris VII, 1990, pp. 21–49.
[20] W. Habicht, Über die Zerlegung strikte definiter Formen in Quadrate, *Comment. Math. Helv.* **12** (1940), 317–322.
[21] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge Univ. Press, Cambridge, 1967.
[22] D. Hilbert, Über die Darstellung definiter Formen als Summe von Formenquadraten, *Math. Ann.* **32** (1888), 342–350.
[23] D. Hilbert, Über ternäre definite Formen, *Acta Math.* **17** (1893), 169–197.
[24] D. Hilbert, Mathematische Probleme, translated by M. W. Newson, *Bull. Amer. Math. Soc.* **8** (1902), 437–479.
[25] D. Hilbert, Beweis für die Darstellbarkeit der ganzen Zahlen durch eine feste Anzahl n-ter Potenzen *(*Waringsches Problem*)*, *Math. Ann.* **67** (1909), 281–300.
[26] D. Hilbert, *Foundations of Geometry*, Open Court Publishing Co., La Salle, Illinois, 1992.
[27] J. R. Joly, Sommes des puissances d-ièmes dans un anneau commutatif, *Acta Arith.* **17**, 37–114.
[28] J.-M. Kantor, Hilbert's problems and their sequels, *Math. Intelligencer* **18** (1996), 21–30.
[29] M. Knebusch, A. Rosenberg, and R. Ware, Structure of Witt rings, quotients of abelian group rings, and orderings of fields, *Bull. Amer. Math. Soc.* **77** (1972), 205–210.
[30] W. Krull, Allgemeine Bewertungstheorie, *J. für die reine und angew. Math.* **167** (1931), 160–196.
[31] T. Y. Lam, *The Algebraic Theory of Quadratic Forms*, W. A. Benjamin, Reading, MA, 1973.
[32] T. Y. Lam, *The theory of ordered fields*, Ring Theory and Algebra III, Proc. Third Oklahoma Conf., 1979, Lecture Notes in Pure and Appl. Math., Vol. 55, Marcel Dekker, New York, 1980, pp. 1–152.
[33] T. Y. Lam, *Orderings, Valuations, and Quadratic Forms*, CBMS Regional Conf. Series in Math., Vol. 52, Amer. Math. Soc., Providence, RI, 1983.
[34] T. Y. Lam, An introduction to real algebra, *Rocky Mtn. J. Math.* **14** (1984), 767–814.
[35] S. Lang, *Algebra*, Addison-Wesley, New York, 1993.
[36] M. Marshall and L. Walter, Signatures of higher level on rings with many units, *Math. Zeit.* **204** (1990), 129–143.
[37] A. Pfister, *Hilbert's 17th Problem and related problems on definite forms*, in **[15]**, 483–489.
[38] G. Pólya, *Über positive Darstellung von Polynomen*, Vierteljschr. Naturforsch. Ges. Zürich, Vol. 73 (1928), 141–145; also in *Collected Papers*, Vol. 2, pp. 309–313, MIT Press, 1974.
[39] V. Powers, Higher level orders on noncommutative rings, *J. Pure and Applied Algebra* **67** (1990), 285–298.
[40] V. Powers, Valuations and higher level orderings in commutative rings, *J. Algebra* **172** (1995), 255–272.
[41] B. Reznick, *Sums of even powers of real linear forms*, Mem. Amer. Math. Soc., Vol. 96, Providence, RI, 1992.
[42] B. Reznick, Uniform denominators in Hilbert's 17th Problem, *Math. Zeit.* **220** (1995), 75–97.

[43] B. Reznick, *Some concrete aspects of Hilbert's 17th Problem*, Séminarie de Structures Algébriques Ordonnées, Vol. 56, Équipe de Logique Mathématique, Université Paris VII, 1996.
[44] P. Ribenboim, 17éme probléme de Hilbert, *Queen's Papers in Pure and Appl. Math.* Vol. 41, 1974, pp. 149–164.
[45] C. Scheiderer, *Real algebra and its applications to geometry in the last 10 years*, Real Algebraic Geometry, Proceedings, Rennes, 1991, Lecture Notes in Math. No. 1524, Springer-Verlag, Berlin, Heidelberg, New York, 1992, pp. 75–96.
[46] C. Sturm, Mémoire sur la résolution des équations numériques, *Ins. France Sc. Math. Phys.* **6** (1835).
[47] T. Szele, On ordered skew fields, *Proc. Amer. Math. Soc.* **3** (1952), 410–413.
[48] A. Tarski, *A decision method for elementary algebra and geometry*, Rand Corporation Publication, Berkeley, CA, 1948.

Department of Mathematics and Computer Science,
Emory University,
Atlanta, GA 30322, USA,
vicki@mathcs.emory.edu

Contributed by Russ Hood, Rio Linda, CA.

NOTES

Edited by **John Duncan**

Computer and Human Reasoning: Single Implicative Axioms for Groups and for Abelian Groups

W. McCune* and A. D. Sands

The search for single axioms for groups has long interested mathematicians. In 1938, Tarski [7] presented the following single equational axiom (in terms of subtraction) for Abelian groups:

$$x - (y - (z - (x - y))) = z, \tag{1}$$

and in 1952, Higman and Neumann [1] presented the following single equational axiom (in terms of division) for ordinary groups:

$$(x/((((x/x)/y)/z)/(((x/x)/x)/z))) = y. \tag{2}$$

We use additive notation, $+, 0, ', -$, for Abelian groups, and multiplicative notation, $\cdot, e, ^{-1}, /$, for ordinary groups. Throughout this note, $-$ and $/$ are binary operations rather than abbreviations for, e.g., $x + y'$ and $x \cdot y^{-1}$.

One might think it trivial, given (2), to obtain a single axiom in terms of product and inverse, by simply rewriting α/β to $\alpha \cdot \beta^{-1}$. Doing so gives a single axiom, but then $\cdot$ is not product, and $^{-1}$ is not inverse. The same situation holds for the Abelian case. Another curious fact is that there is no single equational axiom for groups or for Abelian groups in terms of the three standard operations of product, inverse, and the identity element [8]. Single equational axioms in terms of product and inverse have been reported by Neumann [5] and others [3, 2].

In this note we consider single *implicative* axioms, that is, axioms of the form $\alpha = \beta \Rightarrow \gamma = \delta$. For Abelian groups, an axiom of this type with five variables was given by Sholander [6]. If we allow one of α and β to be a variable, it is trivial to obtain an implicative axiom from an equational one: select any term ζ in the equational axiom, replace it with a new variable v, and add the antecedent $\zeta = v$. Hence we restrict our attention to axioms in which neither α nor β is a variable.

Using a combination of human and computer reasoning—specifically, with the assistance of the automated deduction system Otter [4]—we obtained single implicative axioms for Abelian groups in terms of addition and inverse (4 variables)

*Supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

and in terms of subtraction (4 variables) and axioms for ordinary groups in terms of product and inverse (5 variables) and in terms of division (4 variables).

The axioms of Theorems 2 and 3 were found and proved to be single axioms by the second author, and those of Theorems 1 and 4 were found and proved to be single axioms by Otter [4]. All of the proofs we present are adapted from proofs found (easily) by Otter.

Theorem 1. *Let G be a nonempty set with binary operation $+$ and unary operation $'$ such that for all $x, y, z, u \in G$,*

$$x + y = z + u \Rightarrow (y' + z) + u = x. \tag{3}$$

Then $\langle G; +, ' \rangle$ is an Abelian group.

Proof: We show the existence of an identity element and that the inverse, commutativity, and associativity properties hold for $\langle G; +, ' \rangle$. First, from (3) we obtain

$$(y' + x) + y = x. \tag{4}$$

Then, setting x to $z + w$, and applying (3) again, we have

$$(y' + z) + w = y' + (z + w). \tag{5}$$

(We no longer need (3), i.e., the pair (4) and (5) axiomatizes Abelian groups.)

$$\begin{aligned} a' + a &= ((b' + a') + b) + a && [\text{by } (4)] \\ &= (b' + (a' + b)) + a && [\text{by } (5)] \\ &= b' + ((a' + b) + a) && [\text{by } (5)] \\ &= b' + b. && [\text{by } (4)] \end{aligned}$$

Thus, $a' + a$ is independent of a, and we may call it 0, and from (4), we have that 0 is a left identity: for all x,

$$x' + x = 0, \tag{6}$$

$$0 + x = x. \tag{7}$$

Commutativity:

$$\begin{aligned} a + b &= ((a' + a) + a) + b && [\text{by } (4)] \\ &= ((b' + b) + a) + b && [\text{by } (6)] \\ &= (b' + (b + a)) + b && [\text{by } (5)] \\ &= b + a. && [\text{by } (4)] \end{aligned}$$

Associativity:

$$\begin{aligned} (a + b) + c &= (a + ((c' + b) + c)) + c && [\text{by } (4)] \\ &= (a + (c' + (b + c))) + c && [\text{by } (5)] \\ &= ((c' + (b + c)) + a) + c && [\text{by commutativity}] \\ &= (c' + ((b + c) + a)) + c && [\text{by } (5)] \\ &= (c' + (a + (b + c))) + c && [\text{by commutativity}] \\ &= a + (b + c). && [\text{by } (4)] \end{aligned}$$

Therefore, $\langle G; +, ' \rangle$ is an Abelian group.

Theorem 2. *Let G be a nonempty set with a binary operation $-$ such that for all $x, y, z, u \in G$,*

$$x - y = z - u \Rightarrow u - (z - x) = y. \tag{8}$$

Then $\langle G; - \rangle$ is an Abelian group with $x - y = x + y'$ for all $x, y \in G$.

Proof: First note that (8) holds in Abelian groups when $-$ is interpreted as subtraction. The main part of the proof is to derive Tarski's axiom (1) from (8). By (8) we have

$$x - (y - y) = x = x - (z - z). \tag{9}$$

Applying (8) to (9) gives us $(z - z) - (x - x) = y - y$, which, by itself, yields $u - u = y - y$. Applying (8) to this, we have

$$y - (y - u) = u. \tag{10}$$

Let u be $v - w$, and apply (8) once again to obtain

$$w - (v - y) = y - (v - w). \tag{11}$$

In (10), let y be w and u be $v - y$, and substitute (11) to derive

$$w - (y - (v - w)) = v - y. \tag{12}$$

In (10), let y be v and u be y, and substitute (12) to obtain

$$v - (w - (y - (v - w))) = y,$$

which is Tarski's single axiom (1) for Abelian groups in terms of subtraction.

Theorem 3. *Let G be a nonempty set with binary operation $\cdot$ and unary operation $^{-1}$ such that for all $x, y, z, u, w \in G$,*

$$(x \cdot y) \cdot z = (x \cdot u) \cdot w \Rightarrow u \cdot (w \cdot z^{-1}) = y. \tag{13}$$

Then $\langle G; \cdot, ^{-1} \rangle$ is a group.

Proof: By (13) we have $x \cdot (y \cdot y^{-1}) = x = x \cdot (z \cdot z^{-1})$; hence $(x \cdot (y \cdot y^{-1})) \cdot u = (x \cdot (z \cdot z^{-1})) \cdot u$; hence by (13), $(z \cdot z^{-1}) \cdot (u \cdot u^{-1}) = (y \cdot y^{-1})$; hence $y \cdot y^{-1}$ is independent of y, and we name the element e, which is a right identity,

$$y \cdot y^{-1} = e, \tag{14}$$

$$x \cdot e = x. \tag{15}$$

We have $(e \cdot e) \cdot e = (e \cdot e^{-1}) \cdot e$, and by (13), $e^{-1} \cdot (e \cdot e^{-1}) = e$; hence

$$e^{-1} = e. \tag{16}$$

Next, we have $(e \cdot x) \cdot e = (e \cdot e) \cdot x$, and by (13), $e \cdot (x \cdot e^{-1}) = x$; hence

$$e \cdot x = x. \tag{17}$$

Now, we have $(e \cdot x) \cdot y = (e \cdot (x \cdot y)) \cdot e$, and by (13), $(x \cdot y) \cdot (e \cdot y^{-1}) = x$; hence

$$(x \cdot y) \cdot y^{-1} = x. \tag{18}$$

Finally, by (17) and (18), we have $(e \cdot ((x \cdot y) \cdot z)) \cdot z^{-1} = (e \cdot x) \cdot y$, and by (13), $x \cdot (y \cdot (z^{-1})^{-1}) = (x \cdot y) \cdot z$, and by (18), $x \cdot (((y \cdot z) \cdot z^{-1}) \cdot (z^{-1})^{-1}) = (x \cdot y) \cdot z$, and by (18) again,

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z. \tag{19}$$

Equations (14), (15), and (19) establish the result.

The proof of the following theorem is in the form produced by the theorem prover Otter. The justification $m \to n$ indicates substitution of an instance of the right side of equation m for an instance of a term in the left side of n, and $:i, j, \ldots$ indicates simplification with $i, j, \ldots$. Variables are automatically renamed by the program, and the numbering of the equations reflects the sequence of equations retained by the program.

Theorem 4. *Let G be a nonempty set with a binary operation $/$ such that for all $x, y, z, u \in G$,*

$$x/y = z/u \Rightarrow y/(((z/z)/z)/((x/x)/x)) = u. \tag{20}$$

Then $\langle G; / \rangle$ is a group with $x/y = x \cdot y^{-1}$ for all $x, y \in G$.

Proof: First Otter shows that x/x is independent of x. Terms of the form $(v/v)/v$ are abbreviated as $f(v)$.

2	$x = x$	
3	$x/y = z/u \to y/(f(z)/f(x)) = u$	
5	$x/(f(y)/f(y)) = x$	[3, 2]
6	$(f(x)/f(x))/(f(y)/(((y/z)/(y/z))/(y/z))) = z$	[3, 5]
10	$f(x)/f(x) = f(y)/f(y)$	[5 → 6 : 5, 5, 5]
11	$(f(x)/f(x))/((f(y)/f(y))/f(y)) = f(y)$	[5 → 6 : 5, 5]
19	$(f(x)/f(x))/((((y/y)/(y/y))/(y/y))/((f(z)/f(z))/f(y))) = y$	[10 → 6]
22	$x/(((f(y)/f(y))/f(z))/((f(z)/f(z))/f(z))) = x$	[10 → 5]
25	$(f(x)/f(x))/((f(y)/f(y))/f(z)) = f(z)$	[10 → 11]
29	$(((x/x)/(x/x))/(x/x))/((f(y)/f(y))/f(x)) = f(x)$	[18 → 6 : 5, 5, 19, 19, 25]
31	$(f(x)/f(x))/f(y) = y$	[19 : 29]
36	$x/(y/y) = x$	[22 : 31, 31]
40	$(x/x)/f(y) = y$	[10 → 31 : 31, 31]
54	$x/x = y/y$	[36 → 40 : 36, 36]

We now introduce the element e and use $x/x = e$ to prove the Higman-Neumann single axiom (2). We assert that there are elements A, B, and C for which (2) fails to hold and Otter derives a contradiction.

1	$x = x$	
2	$x/y = z/u \to y/(((z/z)/z)/((x/x)/x)) = u$	
4	$x/x = e$	
5	$A/((((A/A)/B)/C)/(((A/A)/A)/C)) \neq B$	
6	$A/(((e/B)/C)/((e/A)/C)) \neq B$	[5 : 4, 4]
7	$x/y = z/u \to y/((e/z)/(e/x)) = u$	[2 : 4, 4]
9	$x/e = x$	[7, 1 : 4]
10	$e/((e/x)/(e/(x/y))) = y$	[7, 9]
12	$((e/x)/(e/(x/(y/z))))/(e/y) = z$	[7, 10 : 4, 9]
17	$(e/x)/(e/(x/y)) = e/(e/(e/y))$	[10 → 10 : 4]
19	$e/(e/x) = x$	[4 → 10 : 4, 9]
20	$(e/(x/y))/(e/x) = y$	[12 : 17, 19]
24	$(e/x)/((e/y)/(x/(y/z))) = z$	[7, 20 : 19]
28	$(e/((e/x)/y))/x = y$	[19 → 20]
30	$x/((e/y)/((e/x)/(y/z))) = z$	[19 → 24]
103	$x/(((e/y)/z)/((e/x)/z)) = y$	[28 → 30 : 19]
105	$\square$	[103, 6]

Because (20) holds for groups when $/$ is interpreted as division, and because the Higman-Neumann axiom can be derived, the proof is complete.

The main open question remaining is whether there exists a four-variable single implicative axiom in terms of product and inverse for ordinary groups. Also, for the

non-Abelian cases, we do not know whether there exist implicative axioms shorter than ours.

Each of the four proofs we have presented arises from a different amount of translation (by hand) from an Otter-generated proof, and each is in a different style. The proof of Theorem 1 has had the most translation and is probably the most transparent (also it is the easiest theorem). Theorem 4's proof is nearly in the form produced by Otter and is probably opaque to most readers (it certainly is to us). Translating proofs found by computers into human-readable forms is an important problem in automated deduction.

The Otter input files and proofs for these four theorems are available on the World Wide Web through `http://www.mcs.anl.gov/home/mccune/ar/imply/`.

REFERENCES

[1] G. Higman and B. H. Neumann. Groups as groupoids with one law. *Publicationes Mathematicae Debrecen*, 2:215–227, 1952.

[2] K. Kunen. Single axioms for groups. *Journal of Automated Reasoning*, 9(3):291–308, 1992.

[3] W. McCune. Single axioms for groups and Abelian groups with various operations. *Journal of Automated Reasoning*, 10(1):1–13, 1993.

[4] W. McCune. Otter 3.0 Reference Manual and Guide. Tech. Report ANL-94/6, Argonne National Laboratory, Argonne, IL, 1994. See also `http://www.mcs.anl.gov/home/mccune/ar/otter/`.

[5] B. H. Neumann. Another single law for groups. *Bull. Australian Math. Soc.*, 23:81–102, 1981.

[6] M. Sholander, Postulates for commutative groups. *Amer. Math. Monthly*, 66:93–95, 1959.

[7] A. Tarski. Ein Beitrag zur Axiomatik der Abelschen Gruppen. *Fundamenta Mathematicae*, 30:253–256, 1938.

[8] A. Tarski. Equational logic and equational theories of algebras. In K. Schütte, editor, *Contributions to Mathematical Logic*, pages 275–288. North-Holland, Amsterdam, 1968.

Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, Illinois 60439-4844

Department of Mathematics and Computer Science
University of Dundee
Dundee, DD1 4HN
Scotland, U.K.

An Elementary Proof of Horn's Theorem

Roger Webster

In this note we present an elementary proof of Horn's theorem on the intersection of convex sets and point out an error in a classical proof of the result. We discuss the following form of the theorem.

Horn's Theorem. *Let $\mathscr{F}$ be a finite family of compact convex sets in $\mathbb{R}^n$ and let k be a positive integer not exceeding n. Suppose that $\mathscr{F}$ has at least k members and that every k members of $\mathscr{F}$ have a common point. Then each $(n-k)$-flat in $\mathbb{R}^n$ lies in some $(n-k+1)$-flat that meets all members of $\mathscr{F}$.*

Since Horn [3] published his theorem in 1949, it has been extensively researched, with new proofs, generalizations and applications appearing—see Danzer, Grünbaum, Klee [1] and Valentine [5]. Few, if any, of the proofs can be described as *elementary*, in that they are short, simply structured, and depend only upon basic ideas of convexity. Perhaps the one most admirably fulfilling these three conditions is that in Eggleston's classic treatise [2] of 1958, and followed by other authors, for example Lay [4] in 1982. Unfortunately, as we indicate later, this long-standing proof, which occurs in standard texts, contains an oversight that cannot be *immediately* rectified. It was the discovery of this slip that prompted a search for a new elementary proof of Horn's theorem and, in turn, this note.

Both Eggleston [2] and Lay [4] show that the general form of Horn's theorem, stated above, follows easily from the special case $k = n$. Thus to present an elementary proof of the theorem, we need only give an elementary proof of the special case, and this we do in the following lemma. In fact, we prove not the special case itself, but a more illuminating result that provides a surprisingly simple way of finding the line whose existence it guarantees.

Lemma. *Let* $\mathbf{a}$ *be a point of* $\mathbb{R}^n$, *and let* $\mathscr{F}$ *be a finite family* $A_1, \dots, A_r$ *of* r *compact convex sets in* $\mathbb{R}^n$ $(r \geq n)$ *such that every* n *members of* $\mathscr{F}$ *have a common point. Suppose that* L^+ *is a closed halfline in* $\mathbb{R}^n$ *issuing from* $\mathbf{a}$ *that meets a maximum number of members of* $\mathscr{F}$. *Then the line containing* L^+ *meets all members of* $\mathscr{F}$.

Proof: Suppose that the line L containing L^+ does *not* meet all members of $\mathscr{F}$, say it fails to meet the set A_r. Let $A_1, \dots, A_m$ $(n \leq m < r)$ be the members of $\mathscr{F}$ that L^+ does meet. We assume, without loss of generality, that $\mathbf{a}$ does not belong to any member of $\mathscr{F}$.

Let $\mathbf{x}$ be a point on the open halfline $L^+ \setminus \{\mathbf{a}\}$. Since L fails to meet A_r, the point $\mathbf{a}$ does not lie in the compact convex set $\operatorname{conv}(A_r \cup \{\mathbf{x}\})$. Let H be any hyperplane in $\mathbb{R}^n$ *strictly* separating $\mathbf{a}$ and $\operatorname{conv}(A_r \cup \{\mathbf{x}\})$. Then H *strictly* separates $\mathbf{a}$ and A_r, and meets L^+.

Denote by $\pi(A_1), \dots, \pi(A_m), \pi(A_r)$, respectively, the radial projections from $\mathbf{a}$ onto H of $A_1, \dots A_m, A_r$. Then every n of these $m + 1$ convex projections have a common point, so by Helly's theorem applied in H, there is some point, $\mathbf{b}$ say, that belongs to *all* $m + 1$ projections. Thus the closed halfline issuing from $\mathbf{a}$ and passing through $\mathbf{b}$ meets the members of $A_1, \dots, A_m, A_r$ of $\mathscr{F}$. This, however, contradicts the maximality property that characterizes L^+. Hence L *does* meet all members of $\mathscr{F}$. ∎

We conclude by indicating the nature of the error in Eggleston's proof [2] of the special case $k = n$ of Horn's theorem. His proof is by induction on the number of sets in the finite family $\mathscr{F}$. During the induction step of his argument, however, he assumes the existence of a *halfline* issuing from a given point and meeting a certain number of members of $\mathscr{F}$, whereas his induction hypothesis *only* guarantees the existence of a *line* through the point meeting the requisite number of members of $\mathscr{F}$, and this is not sufficient for the proof to continue.

REFERENCES

1. L. W. Danzer, B. Grünbaum and V. Klee, Helly's theorem and its relatives, *Amer. Math. Soc. Symposium on Convexity, Proceedings of Symposia in Pure Mathematics* 7 (1963), 101–180.
2. H. G. Eggleston, *Convexity*, Cambridge University Press, 1958.

3. A. Horn, Some generalizations of Helly's theorem on convex sets, *Bull. Amer. Math. Soc.* **55** (1949), 923–929.
4. S. R. Lay, *Convex sets and their applications*, Wiley, New York, 1982.
5. F. A. Valentine, *Convex sets*, McGraw-Hill, New York, 1964.

Pure Mathematics Section
School of Mathematics and Statistics
University of Sheffield
Sheffield S3 7RH
England
r.j.webster@sheffield.ac.uk

Packability of Five Spheres on a Sphere Implies Packability of Six

James Angelos, George Grossman, Yury Ionin, Edwin Kaufman, Terry Lenker, and Leela Rakesh

While trying to determine how many nonoverlapping spheres of radius r_2 can be tangent to (and outside of) a sphere of radius r_1, we discovered a startling fact: If r_1 and r_2 are such that it is not possible to pack (i.e., place) six spheres as described, then it is not possible to pack five spheres either; the maximum number for which one can hope drops from six to four, skipping five, as r_2/r_1 increases from $1 + \sqrt{2}$. Stated alternately, whenever r_1 and r_2 are such that five spheres can be packed, then it is always possible to rearrange the packing to make room for a sixth sphere. We believe that this result is of interest because it can be understood by people with little mathematical background, and at the same time it illustrates how counterintuitive mathematics can be at times. The proof also nicely illustrates the usefulness of using different coordinate systems, among other things. The idea of packing spheres around another sphere has its beginnings from an apparent conversation between Isaac Newton and David Gregory in 1694. The question that arose between them was: "Can a rigid material sphere be brought into contact with 13 other spheres of the same size?" Gregory thought the answer was yes, while Newton thought no. 180 years later in 1874, Newton's answer was shown to be the case [**1**]. Our results follow from the following two-part theorem.

Theorem. *Let r_1 and r_2 be positive real numbers.*

(1) *If $r_2/r_1 \le 1 + \sqrt{2}$, then it is possible to place six nonoverlapping spheres of radius r_2 tangent to (and outside of) a sphere of radius r_1.*
(2) *If $r_2/r_1 > 1 + \sqrt{2}$, then it is not possible to place five nonoverlapping spheres of radius r_2 tangent to (and outside of) a sphere of radius r_1.*

Proof: (1) This part of the theorem will follow once we have displayed a packing for $r_2/r_1 = 1 + \sqrt{2}$, since decreasing r_2/r_1 makes it easier to pack the spheres. Without loss of generality, suppose $r_1 = 1$ and $r_2 = 1 + \sqrt{2}$. Take the origin to be

at the center of the sphere of radius 1, and place the centers of the six spheres of radius $1 + \sqrt{2}$ as $(\pm(2 + \sqrt{2}), 0, 0)$, $(0, \pm(2 + \sqrt{2}), 0)$, and $(0, 0, \pm(2 + \sqrt{2}))$. Then the distance from the center of each of these six spheres to the origin is $2 + \sqrt{2} = 1 + (1 + \sqrt{2})$, so these spheres are tangent to the sphere of radius 1. Furthermore, the distance between the centers of any two of the spheres of radius $1 + \sqrt{2}$ is either $2(2 + \sqrt{2}) > 1 + \sqrt{2} + 1 + \sqrt{2}$, or else by the distance formula it is $\sqrt{(2 + \sqrt{2})^2 + (2 + \sqrt{2})^2} = \sqrt{12 + 8\sqrt{2}} = 2 + 2\sqrt{2} = 1 + \sqrt{2} + 1 + \sqrt{2}$, so these spheres do not overlap and part (1) of the theorem is proved. Figure 1 shows this packing.

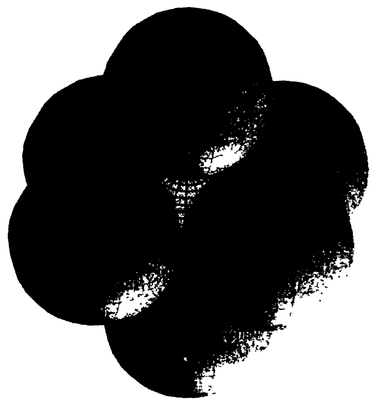


Figure 1

(2) Suppose (by way of contradiction) that $r_2/r_1 > 1 + \sqrt{2}$, but five nonoverlapping spheres of radius r_2 can be placed tangent to (and outside of) a sphere of radius r_1. Consider a plane passing through the center of the sphere of radius r_1 and the centers of any two of the spheres of radius r_2, let the intersection of this plane with the sphere of radius r_1 be called the equator, and let the two points on the sphere farthest from the equator be called the north pole and the south pole. Then by the pigeonhole principle, either at least two of the remaining three spheres of radius r_2 must touch the sphere of radius r_1 in the northern hemisphere (or on the equator), or else at least two of these three spheres must touch the sphere of radius r_1 in the southern hemisphere (or on the equator). Without loss of generality, assume the former; then the points of contact of at least four of the spheres of radius r_2 with the sphere of radius r_1 will be in the northern hemisphere (or on the equator). Now take the origin to be at the center of the sphere of radius r_1 and the positive z-axis to contain the north pole. In spherical coordinates let the points of contact of four spheres in the northern hemisphere (or on the equator) with the sphere of radius r_1 be (ρ, θ_1, ϕ_1), (ρ, θ_2, ϕ_2), (ρ, θ_3, ϕ_3), and (ρ, θ_4, ϕ_4), where $\rho = r_1$, $0 \le \theta_i < 2\pi$ for $i = 1, 2, 3, 4$, and $0 \le \phi_i \le \pi/2$ for $i = 1, 2, 3, 4$. Now there must be indices l and k with $|\theta_l - \theta_k| \le \pi/2$, or $|\theta_l - \theta_k - 2\pi| \le \pi/2$. Let us rotate the coordinate system about the z-axis so that θ_k becomes zero. Then the contact points of spheres l and k with the sphere of radius r_1 have spherical coordinates (ρ, θ, ϕ_l) and $(\rho, 0, \phi_k)$ respectively, where we can take $|\theta| \le \pi/2$ and $0 \le \phi_l, \phi_k \le \pi/2$. Now consider the (smaller) angle α between the vectors from the origin to these contact points; returning to rectangu-

lar coordinates, we have

$$\cos\alpha = [\rho\cos\theta\sin\phi_l, \rho\sin\theta\sin\phi_l, \rho\cos\phi_l]\cdot[\rho\sin\phi_k, 0, \rho\cos\phi_k]/(\rho\cdot\rho)$$
$$= \cos\theta\sin\phi_l\sin\phi_k + \cos\phi_l\cos\phi_k \geq 0,$$

so $\alpha \leq \pi/2$. Now consider the intersection of spheres l and k and the sphere of radius r_1 with a plane which passes through their centers (see Figure 2); since

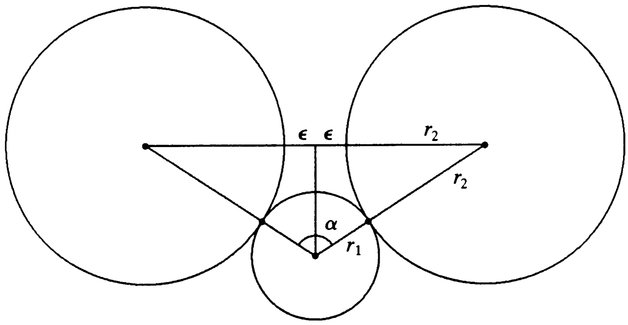


Figure 2

spheres l and k are assumed not to overlap, we have $\varepsilon \geq 0$. Thus

$$\frac{r_2}{r_1 + r_2} \leq \frac{r_2 + \varepsilon}{r_1 + r_2} = \sin\frac{\alpha}{2} \leq \sin\frac{\pi}{4} = \frac{1}{\sqrt{2}},$$

so $\sqrt{2}r_2 \leq r_1 + r_2$, so $r_2(\sqrt{2} - 1) \leq r_1$, so

$$\frac{r_2}{r_1} \leq \frac{1}{\sqrt{2} - 1} = 1 + \sqrt{2}.$$

This contradicts the hypothesis that $r_2/r_1 > 1 + \sqrt{2}$, and part (2) of the theorem is prove. ■

ACKNOWLEDGMENT. The last-named author is grateful to Drs. Donald Tomalia and Marc Mansfield for originating this problem in connection with dendrimer polymer applications during the author's sabbatical leave at the Michigan Molecular Institute.

REFERENCES

[1] H. S. M. Coxeter, An Upper Bound for the Number of Equal Nonoverlapping Spheres that can Touch Another of the Same Size, Convexity, *Proc. of Symp. in Pure Math.*, American Mathematical Society, Volume 7, Ed. Victor Klee, 1963.

Department of Mathematics,
Central Michigan University,
Mt. Pleasant, MI 48859
james.angelos@cmich.edu

UNSOLVED PROBLEMS

Edited by: **Richard Nowakowski & Richard Guy**

In this department the MONTHLY presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Nowakowski, Department of Mathematics & Statistics & Computing Science, Dalhousie University, Halifax N.S., Canada B3H 3J5.

What's the Bound on the Average Number of Normals?

Kathy Hann

The Problem. The question, what are the bounds on the average number of normals through a point in a convex body?, was considered by Santaló in 1944 **[15]**. This question has been answered for many special cases and it has been shown in Euclidean spaces that the average is finite and bounded above by 12 in the plane and by 62 in 3-space. See **[10]** and **[14]**. But numerical data suggest that these numbers are too high; indeed, it may be that the strict upper bound is 8 in the plane and 26 in 3-space. The main unsolved problem we propose is

1. *What is the strict upper bound on the average number of normals through a point in a convex body and what types of bodies attain that upper bound?*

Notation and Definitions. Let K be a convex body in R^m, $m \geq 2$, so K is a compact, convex set with interior points. A *normal* to K at a point x in the boundary of K, ∂K, is a ray with endpoint at x, perpendicular to a support plane H of K at x, and contained in the halfspace bounded by H that contains K. In the example in Figure 1, K has an infinite number of normals at point A and the

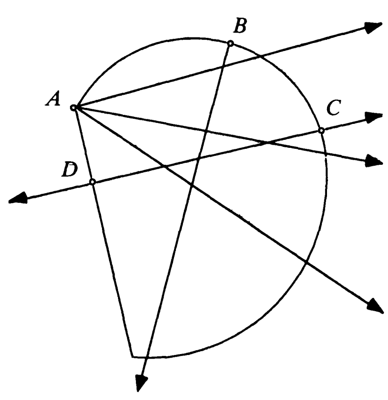


Figure 1. Some normals to a convex body.

normals emanating from C and D are distinct even though they lie on the same line.

For each point p in K let $n(K, p)$ be the number of normals to K passing through p. Let $V(K)$ be the m-dimensional Lebesgue measure (volume) of K,

$$I(K) = \int_K n(K, p)\, dV$$

and

$$n(K) = \frac{I(K)}{V(K)}.$$

Then $n(K)$ is the average of $n(K, p)$ over K.

For a simple example, consider a square Q. Each interior point of Q has 8 normals passing through it. Thus $n(Q, p) = 8$ almost everywhere and hence $n(Q) = 8$.

Notice that every point p in the interior of K has at least two normals passing through it, one from the point on the boundary of K furthest from p and one from the point on the boundary closest to p, so we have the lower bound, $n(K) \geq 2$. This bound is attained for any ball B.

History. Normals have been studied for a very long time. Apollonius of Perga was probably the first person to write down a rigorous discussion of maxima and minima, as he called them. He wrote the *Conics* **[1]** sometime near the year 200 B.C. In this work he demonstrated, indirectly, that the number of maxima and minima through a point in an ellipse is always 2 or 4 and gave conditions that determine this number. Using these conditions, one can see that the number is 2 outside the evolute of the ellipse and is 4 inside the evolute; see [**1**, pp. 177–8].

Since then a number of people have studied local properties of our function $n(K, p)$; see **[4]**, **[6]**, [7], **[11]**, **[12]**, **[13]**, **[16]**, and **[17]**.

For sets of constant width, each diameter is a double normal. Among others, Hammer **[9]**, Besicovitch & Zamfirescu **[3]**, and Bárány & Zamfirescu **[2]** have investigated local properties of diameters.

Known Bounds. Chakerian **[5]** completely solved our problem for planar convex bodies of constant width. Specifically he showed that in this case

$$2 \leq n(K) \leq \frac{2\pi}{\pi - \sqrt{3}}$$

with equality holding in the upper bound only for a Reuleaux triangle and in the lower bound only for a circular disk.

Hann **[10]** proved the following results for polytopes and bodies with sufficiently smooth boundaries. In the plane we have

$n(K) \leq 12$

$n(K) \leq 8$ for K centrally symmetric, and

$n(K) \leq 6$ if all the centers of curvature of K are inside K.

Equality holds in the centrally symmetric case for all regular *2m-gons* K. In 3-space we have

$n(K) \leq 62$ and

$n(K) \leq 26$ for K centrally symmetric.

The second inequality is sharp since equality holds for cuboids. Recently, Hug **[14]** proved that these bounds hold for arbitrary convex bodies.

These results generalize to $\mathbb{R}^m$, $m \geq 3$. Let

$$DK = K + (-K) = \{x_1 - x_2 : x_1, x_2 \in K\}$$

be the difference body of K. Then we have

$$n(K) \leq \frac{V(K + DK)}{V(K)} - 1 \quad \text{and}$$

$$n(K) \leq 3^m - 1 \text{ if } K \text{ is centrally symmetric.}$$

The second inequality is sharp since equality holds for an m-cube.

Finally, Hug **[14]** characterized those centrally symmetric $2m$-gons whose average is 8 as those with the property that the convex hull of opposite edges is a rectangle. These are the centrally symmetric polygons inscribed in a circle, or the blocklike polygons. One can easily generalize his argument to show that the centrally symmetric polytopes in $\mathbb{R}^m$ whose average is $3^m - 1$ are the blocklike polytopes, that is, the centrally symmetric polytopes inscribed in an m-sphere.

Hug **[14]** also proved for a triangle T that $4 < n(T) \leq 6$, $n(T) = 6$ if and only if T is not obtuse, and that every value in $(4, 6]$ is attained by some triangle.

Normals to Polytopes and an Euler-type Identity. For a polytope P we can calculate $n(P)$ by evaluating a finite number of areas, namely the area of P and the areas of all the k-wedges of P, which we now define.

Consider a polytope P in $\mathbb{R}^m$. For each point $x \in \partial P$, let the normal chords corresponding to x be the chords of P that lie in a normal of P that emanates from x. Let F be a fixed i-dimensional face of P and $k = m - i$. Then the k-wedge of P corresponding to F is the convex subset of P containing F and all the normal chords of P that are also normal to the affine hull of F. The 1-wedge at edge e and the 2-wedge at vertex v of a polygon are pictured below.

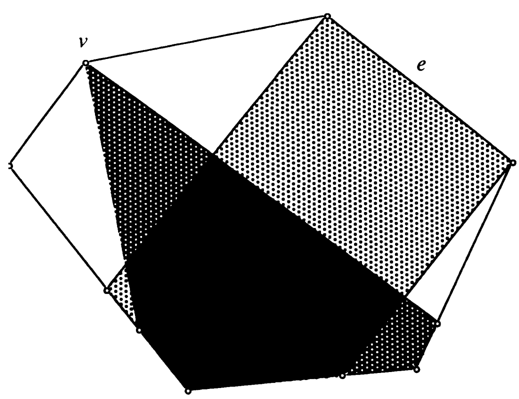


Figure 2. A 1-wedge and a 2-wedge of a polygon.

For each point $p \in P$, $n(P, p)$ is the number of k-wedges that contain p, so one way to calculate $n(P)$ is to add up all the areas of the k-wedges of P and divide by the area of P.

There is an elementary geometric argument, see **[10]**, which shows that for a polygon P the sum of the areas of all the 1-wedges equals the sum of the areas of

all the 2-wedges. Thus, the sum of the areas of the wedges corresponding to edges equals the sum of the areas of the wedges corresponding to vertices. This analog of Euler's relation generalizes to $\mathbb{R}^m$. In $\mathbb{R}^3$ the generalization is that the sum of the volumes of all the 3-wedges minus the sum of the volumes of all the 2-wedges plus the sum of the volumes of all the 1-wedges equals twice the volume of the polytope. In $\mathbb{R}^m$ the alternating sum adds up to $1-(-1)^m$ times the volume of the polytope. The only published proofs of these relations involve high-powered geometric analysis; see **[10]** and **[14]**.

The next problem we propose is

2. *Find an elementary geometric proof in $\mathbb{R}^3$ that the sum of the volumes of all the wedges corresponding to the vertices minus the sum of the volumes of all the wedges corresponding to the edges plus the sum of the volumes of all the wedges corresponding to the facets of a polytope equals twice the volume of the polytope. Generalize the argument to $\mathbb{R}^m$.*

Other Related Problems. 3. *What is the strict lower bound on the average number of normals through a point in a convex polytope?*

4. *What is the strict upper bound on the average number of normals through a point in a convex body whose boundary has no flat parts, that is all (or almost all) of the principal curvatures are nonzero?*

5. *Generalize the solutions for problems* 1–4 *to a Minkowski space.*

REFERENCES

1. Apollonius (Heath), *Apollonius of Perga, Treatise on Conic Sections. Edited in modern notation with introductions including an essay on the earlier history of the subject by T.L. Heath*, M. A. Cambridge, (1896).
2. I. Bárány & T. Zamfirescu, Diameters in typical convex bodies, *Canad. J. Math.*, 42 (1990), 50–61.
3. A. S. Besicovitch & T. Zamfirescu, On pencils of diameters in convex bodies, *Rev. Roumaine Math. Pures Appl.* 11 (1966), 637–639.
4. R. C. Bose & S. N. Roy, A note on the arc centroid of a closed convex oval, *Bull. Calcutta Math. Soc.* 27 (1935), 111–118.
5. G. D. Chakerian, The number of diameters through a point inside an oval, *Unión Matemática Argentina* 29 (1984), 282–290.
6. G. D. Chakerian & S. K. Stein, On the centroid of a homogeneous wire, *Mich. Math. J.* 11 (1964), 189–192.
7. V. Douglas, E. Heil, and C. Lübbert, Geradenfamilien, Enveloppen und Evoluten, *J. Reine Angew. Math.* 77 (1976), 370–383.
8. C. Godbersen, *Der Satz vom Vektorbereich in Räumen beliebiger Dimension*, Dissertation, Göttingen, (1938).
9. P. C. Hammer, Diameters of convex bodies, *Proc. Amer. Math. Soc.* 5 (1954), 304–306.
10. K. Hann, The average number of normals through a point in a convex body and a related Euler-type identity, *Geom. Dedicata* 48 (1993), 27–55.
11. T. Hayashi, Some geometrical applications of the Fourier series, *Rend. Circ. Mat. Palermo* 50 (1926), 96–102.
12. E. Heil, Concurrent normals and critical points under weak smoothness assumptions, *Ann. New York Acad. Sci.* 440 (1985), 170–178.
13. E. Heil, Existenz eines 6-Normalenpunktes in einem konvexen Körper, *Arch. Math.* 32 (1979), 412–416.
14. Daniel Hug, On the mean number of normals through a point in the interior of a convex body, *Geom. Dedicata* 55 (1995), 319–340.
15. L. A. Santaló, Note on convex spherical curves, *Bull. Amer. Math. Soc.* 50 (1944), 528–534.
16. A. Schatteman & G. Valette, On the Normals of Closed Plane Curves, *Simon Stevin* 59 (1985), 235–245.
17. H. Tietz, Über die Anzahl der stabilen Ruhelagen eines Würfels, *Elem. Math.* 3 (1948), 97–100.

THE AUTHORS

SHAWNEE L. MCMURRAN earned her Ph.D. at the University of California at Riverside in 1991. Her background is in partial differential equations and her current research interests include women and mathematics at Cambridge.

JAMES J. TATTERSALL earned his Ph.D. at the University of Oklahoma in 1971 and has been on the faculty of Providence College ever since. He has been a visiting scholar at Wolfson College, Cambridge, a visiting mathematician at the American Mathematical Society, and a visiting professor of mathematics at the United States Military Academy at West Point. He is the Historian/Archivist of the Northeastern Section of the Mathematical Association of America. His current research interests include women and mathematics at Cambridge and the Lucasian Professorship at Cambridge. His hobbies include gardening and reading.

ROMAN KOSSAK received his Ph.D. from Warsaw University in 1982 and his Habilitated Doctor degree from the Polish Academy of Sciences in 1992. His research interests include model theory and logical foundations of arithmetic. He has been teaching mathematics in the U.S. since 1986. He has held visiting positions at Baruch College/CUNY and University of Connecticut at Storrs. He teaches at Bronx Community College/CUNY in New York.

JOHN BRILLHART was born in Berkeley, educated in the city schools, and received his PhD at Berkeley in 1967. A student of D. H. Lehmer and R. M. Robinson, his first research was in computational number theory in 1957 using the IBM 701. An mathematics professor at the University of Arizona since 1967, he has recently worked with Richard Blecksmith and the late Irving Gerst on triple and quintuple product identities, and with J. S. Lomont on elliptic integrals and orthogonal polynomials. Exploring the beauties of mathematics and developing the subject as a craft have been lifelong preoccupations.

PATRICK MORTON is an associate professor at Wellesley College. He obtained his PhD at Michigan under the guidance of Donald J. Lewis, and his BA at the University of Arizona, where he studied with John Brillhart. He has held postdoctoral positions at Caltech and Harvard, and was a von Humboldt Fellow at the University of Heidelberg. He likes to muse on questions in the intersection of number theory and dynamical systems, and on the creative process in mathematics. He is co-author of an integrated middle school curriculum in mathematics and social studies, and struggles to learn classical guitar.

EDWARD F. ABOUFADEL first decided to become a mathematics professor after taking a series of honors calculus courses from John Hocking at Michigan State University. In 1992, he received his doctorate from Rutgers University, studying differential equations under Jane Cronin Scanlon. His mathematical interests are best described as a mix of applied mathematics and collegiate mathematics education, and he would like someday to teach a course in wavelets and fingerprint compression to undergraduates. He is in a tenure-track position at Grand Valley State University. Ed has been known to keep a diary.

VICTORIA POWERS received an A.B. from the University of Chicago in 1980 and a Ph.D from Cornell University in 1985. Her Ph.D advisor was Alex F. T. W. Rosenberg, which makes her a direct (mathematical) descendant of David Hilbert. She taught at the University of Hawaii for two years and has been at Emory University since 1987. Her mathematical interests include quadratic forms, ordered fields, and algebraic geometry, particularly computational aspects of real algebraic geometry. She credits her great-aunt Sums of Even with sparking her interest in higher-level orderings.

WILLIAM MASSEY was born in central Illinois and grew up there during the great depression. He obtained his bachelor's and master's degree from the University of Chicago just before World War II, and served in the U.S. Navy for four years during the war. After the war he got his Ph.D. at Princeton. After ten years on the faculty at Brown University and thirty one years at Yale, he retired with the title Professor Emeritus in 1991. His four books and most of his research papers have been concerned with algebraic topology. Retirement has permitted him to spend more time on his favorite hobby, bird watching.

PROBLEMS AND SOLUTIONS

Edited by:
Richard T. Bumby, Fred Kochman and Douglas B. West

Proposed problems should be sent to the MONTHLY PROBLEMS address given on the inside front cover. Please include solutions and relevant references. Two copies of all items needed to evaluate the problem should be sent. A third copy of the the problem and solution is often useful; please include one if possible.

Solutions of published problems should arrive at the MONTHLY PROBLEMS address given on the inside front cover before May 31, 1997. If possible, solutions should be typed with double spacing. Two copies suffice. Several solutions may be mailed together, but they should be on separate sheets of paper. The problem number and the solver's name and mailing address should appear on each solution. A mailing label should be included if an acknowledgment is desired.

The published solution is likely to be based on a solution that is complete and correct. Additional information, such as references to other appearances of the problem or its solution, is also welcome.

An asterisk (*) after the number of a problem, or part of a problem, indicates that no solution is currently available.

PROBLEMS

10557. *Proposed by Nick MacKinnon, Winchester College, Winchester, U. K..*

Naismith's rule allows walkers to compute the time for their journeys. The time is given by allowing a walking speed of 4 km/hr, but adding an extra minute for each 10m of ascent. A conical mountain has base radius 1650m and vertical height 520m. Points A and B are diametrically opposite at the base of the mountain. How should a path be constructed between A and B on the surface of the mountain which minimizes the time taken to walk from A to B?

10558. *Proposed by Zhang Chengyu, Hubei University, Wuhan, China.*

Let p be a prime and let k be a positive integer. Let $a_1, a_2, \ldots, a_{p^k}$ be any p^k integers. We define the *adjustment* of these integers to be the p^k integers $b_1, b_2, \ldots, b_{p^k}$, where $b_j = a_{j+1} + a_{j+2} + \cdots + a_{j+p}$ interpreting subscripts modulo p^k. For example, if $p = 2$ and $k = 2$, one adjustment of 1, 1, 3, 4 gives 4, 7, 5, 2. Prove that after p^k adjustments of $a_1, a_2, \ldots, a_{p^k}$, the list consists entirely of integers divisible by p.

10559. *Proposed by Michael Golomb, Purdue University, West Lafayette, IN.*

Determine the class $\mathcal{U}$ of real-valued differentiable functions that satisfy the functional equation

$$u(2x) = 2u(x)u'(x)$$

for all real x and that are real analytic near $x = 0$.

10560. *Proposed by Emre Alkan (student), Bosphorus University, İstanbul, Turkey.*

Consider a convex quadrilateral $ABCD$, and choose points P, Q, R, and S on sides AB, BC, CD, and DA, respectively, with

$$\frac{|PA|}{|PB|} = \frac{|RD|}{|RC|} \text{ and } \frac{|QB|}{|QC|} = \frac{|SA|}{|SD|}.$$

Let K denote the area of $ABCD$, and let K_A, K_B, K_C, and K_D denote the areas of SAP, PBQ, QCR, and RDS, respectively. Show that $K^4 \geq 2^{12} K_A K_B K_C K_D$ and determine a necessary and sufficient condition for equality.

10561. *Proposed by Jean Anglesio, Garches, France.*

Show that

$$\gamma = \lim_{u \to \infty} \int_{1/u}^{u} \left(\frac{1}{2} - \cos x \right) \frac{dx}{x}$$

and

$$\gamma = \lim_{u \to \infty} \int_{1/u}^{u} \left(\frac{1}{1+x} - \frac{\cos x}{x} \right) \frac{dx}{x},$$

where γ denotes Euler's constant $\lim_{n \to \infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n \right)$.

10562. *Proposed by K. Hinderer & M. Stieglitz, University of Karlsruhe, Karlsruhe, Germany.*

Suppose K is a fixed positive integer greater than 1.

(a) If $2 \leq n \leq K$, how large is the maximal product W_n of *exactly n* positive integers whose sum equals K?

(b) If $2 \leq n \leq K$, how large is the maximal product V_n of *at most n* positive integers whose sum equals K?

10563. *Proposed by Harold G. Diamond, University of Illinois, Urbana, IL.*

Let F be locally integrable on $[0, \infty)$ with Laplace transform $f(s) = \int_0^\infty e^{-sx} F(x)\, dx$. It is easy to show that if $\left(1/x\right) \int_0^x F(t)\, dt$ has a limit as $x \to \infty$ (the *average value* of F), then $sf(s)$ converges to the same limit as $s \to 0^+$.

Let

$$F(x) = \begin{cases} 1 & \text{for } 0 \leq x < 1 \\ (-1)^{n+1} & \text{for } 2^n \leq x < 2^{n+1}\ (n \geq 0). \end{cases}$$

Clearly, F is bounded from below and has no average value, from which it can be shown that $sf(s)$ has no limit at $s = 0$. Determine the asymptotic behavior of $sf(s)$ as $s \to 0^+$.

NOTES

(**10562**) The case $n = K$ of (**b**) appears as Problem 15 in D. J. Newman, *A Problem Seminar*, Springer, 1982. (**10563**) With the other assumptions about $F(x)$, if $\lim_{s\to 0^+} sf(s) = A$, the Laplace transform version of the Hardy-Littlewood-Karamata tauberian theorem would imply that A is the average value of $F(x)$. See D. V. Widder, *The Laplace Transform*, Princeton, 1941 (especially Theorem 4.6 of chapter V) for more details.

SOLUTIONS

Damped Oscillation

10250 [1992, 782]. *Proposed by Xin Li, University of Central Florida, Orlando, FL.*

Assume that $k \in \mathbb{Z}, k > 1$, and $\lambda \in \mathbb{R}, \lambda > 0$. Define

$$S(t) = \sin kt + \lambda \sin(k-1)t$$

and let $\langle t_i \rangle$ with $0 < t_1 < t_2 < \ldots < \pi$ be all zeros of $S'(t)$ in the interval $(0, \pi)$.

Show that $|S(t_i)| > |S(t_{i+1})|$ for all i, i.e. that the sequence of relative maxima of $|S(t)|$ on this interval is strictly decreasing.

Solution by the proposer. Define

$$C(t) = \cos kt + \lambda \cos(k-1)t$$

and

$$f(t) = \left| e^{ikt} + \lambda e^{i(k-1)t} \right| = \left(1 + \lambda^2 + 2\lambda \cos t \right)^{1/2}.$$

We observe that $f(t)$ is strictly decreasing for $0 \le t \le \pi$. Also,

$$|S(t)| \le \left(|S(t)|^2 + |C(t)|^2 \right)^{1/2} = f(t).$$

Note that, with

$$\alpha_j = \frac{2j-1}{2k}\pi \text{ and } \beta_j = \frac{2j-1}{2(k-1)}\pi,$$

we have

$$\operatorname{sgn} S'(\alpha_j) = -\operatorname{sgn} S'(\beta_j) = (-1)^{j+1}$$

and

$$\operatorname{sgn} C(\alpha_j) = -\operatorname{sgn} C(\beta_j) = (-1)^{j+1}$$

for $j = 1, 2, \ldots, k-1$. So for each j, there exist $t_j, s_j \in (\alpha_j, \beta_j)$ such that

$$S'(t_j) = 0 \text{ and } C(s_j) = 0.$$

Furthermore, $S'(t)$ cannot have more than k zeros in $(0, \pi)$, so it has at most one zero other than the t_j. By considering the signs at the endpoints, we infer that $S'(t)$ has no zero in $[0, \pi/(2k)]$ or in $[\beta_j, \alpha_{j+1}]$ for $j = 1, 2, \ldots, k-2$. Since

$$\begin{aligned}\operatorname{sgn} S'(\beta_{k-1}) \cdot \operatorname{sgn} S'(\pi) &= (-1)^{k-1} \cdot \left(k(-1)^k + \lambda(k-1)(-1)^{k-1} \right) \\ &= -k + \lambda(k-1),\end{aligned}$$

$S'(t)$ has one more simple zero, which we denote t_k, which will be located in (β_{k-1}, π), if and only if $\lambda < k/(k-1)$.

For $j = 1, 2, \ldots, k-1$, t_j is the location of the maximum of $|S(t)|$ on $[\alpha_j, \beta_j]$, while $S(s_j) = f(s_j)$. Hence,

$$f(t_j) \geq \left|S(t_j)\right| \geq \left|S(s_j)\right| = f(s_j) > f(t_{j+1}) \geq \left|S(t_{j+1})\right|,$$

since $s_j < t_{j+1}$ and $f(t)$ is strictly decreasing. (Note that this also shows that each $t_j \leq s_j$.)

Solved also by M. Bowron and A. D. Melas (Greece).

A Recurrence of Fibonacci

10316 [1993, 589]. *Proposed by Richard K. Guy, University of Calgary, Calgary, Alberta, Canada, and Richard J. Nowakowski, Dalhousie University, Halifax, N.S., Canada.*

For what pairs of integers a, b, does ab exactly divide $a^2 + b^2 + 1$?

Solution by Daniel Schepler (student), Washington University, St. Louis, MO. The problem is unchanged by the interchange of a and b and by negation, so we may assume that $a \geq b > 0$. We prove that $(a, b) = (F_{2n+1}, F_{2n-1})$ for nonnegative n, where F_n denotes the nth Fibonacci number with $F_{-1} = 1$ and $F_0 = 0$.

By the well-known identity $F_{2n-1}F_{2n+1} = F_{2n}^2 + 1$, which can be proved by induction, we have

$$F_{2n+1}^2 + F_{2n-1}^2 - 2F_{2n+1}F_{2n-1} = (F_{2n+1} - F_{2n-1})^2 = F_{2n}^2 = F_{2n+1}F_{2n-1} - 1,$$

which yields $F_{2n+1}^2 + F_{2n-1}^2 + 1 = 3F_{2n+1}F_{2n-1}$. Therefore, (F_{2n+1}, F_{2n-1}) is a solution.

To prove that we have all solutions, we use induction on a. If $a = b$, then $a^2|(2a^2 + 1)$ requires $a = b = 1$; this is the solution above for $n = 0$. Hence we may assume that $a > b$. Suppose $a^2 + b^2 + 1 = kab$. Let $a' = b$ and $b' = kb - a = (b^2 + 1)/a$. Straightforward computation yields

$$(a')^2 + (b')^2 + 1 = kb(kb - a) = ka'b',$$

and hence (a', b') is a solution.

After showing that $0 \leq b' \leq a' < a$, we can apply the induction hypothesis. Both a' and b' are positive. Also $b' = (b^2+1)/a < (b^2+1)/b \leq b+1$. Since b' is an integer, we have $b' \leq b = a'$. Also $a' = b < a$. Hence the induction hypothesis yields $a' = b = F_{2n+1}$ and $b' = kb - a = F_{2n-1}$ for some nonnegative integer n. Furthermore, we have computed in this case that $k = 3$. Therefore,

$$a = kb - b' = 3F_{2n+1} - F_{2n-1} = 2F_{2n+1} + F_{2n} = F_{2n+1} + F_{2n+2} = F_{2n+3}.$$

Thus $(a, b) = (F_{2n+3}, F_{2n+1})$ has the desired form, which completes the proof.

Editorial comment. Allan Pedersen, Michael Vowe, J. H. Steelman, David Doster, and J. G. Mauldon noted that ab divides a^2+b^2+1 if and only if ab divides $(a^2+1)(b^2+1)$, which holds if and only if $a|(b^2+1)$ and $b|(a^2+1)$, and hence this problem is the same as problem E3210 (1987, 457; 1988, 877). Emre Alkan, John Christopher, J. P. Robertson, Kiran Kedlaya, David Doster, Man-Keung Siu, and Charles Lanski noted that the problem is also equivalent to problem 10203 (1992, 265; 1994, 279). M. J. Knight and the Chico Problem Group found the equation treated in L. J. Mordell, *Diophantine Equations*, Academic Press, 1966, Section 20, page 299. D. E. Manes reported that D. O. Shklarsky, N. N. Chentzov, and I. M. Yaglom, *The USSR Olympiad Problem Book*, W.H. Freeman, 1962, pp. 205-208 presents an algorithm for solving $x^2 + y^2 + z^2 = kxyz$.

Solved by 85 readers, including those cited, and the proposer. Two incomplete solutions and three incorrect solutions were received.

Closely Related Triangles

10317 [1993, 590]. *Proposed by Juan Bosco Romero Márquez, Universidad de Valladolid, Valladolid, Spain.*

Let $\triangle ABC$ be inscribed in a circle $\mathcal{C}$ and let A', B', C' be the midpoints of the arcs $\overset{\frown}{BC}$, $\overset{\frown}{CA}$, $\overset{\frown}{AB}$, respectively.

(**a**) Prove that the incenter of $\triangle ABC$ is the orthocenter of $\triangle A'B'C'$.

(**b**) Prove that the pedal triangle of $\triangle A'B'C'$ is homothetic to $\triangle ABC$

Solution by O. P. Lossers, University of Technology, Eindhoven, The Netherlands. Referring to Figure 10317 we use ancient arguments in standard Euclidean geometry as follows:

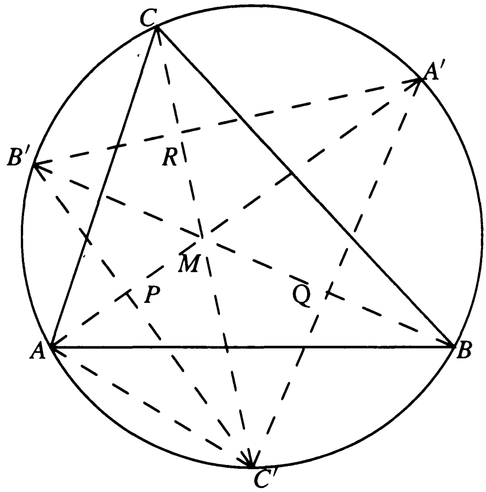


Figure 10317

(**a**)$\angle A'AC = \frac{1}{2}\overset{\frown}{A'C} = \frac{1}{2}\overset{\frown}{BA'} = \angle BAA'$. So AA' is a bisector in $\triangle ABC$. On the other hand, suppose that AA' and $B'C'$ intersect in the point P. Then

$$\begin{aligned}\angle C'PA' &= \frac{1}{2}\left\{\left(\overset{\frown}{C'B} + \overset{\frown}{BA'}\right) + \overset{\frown}{B'A}\right\}\\ &= \frac{1}{4}\left\{\overset{\frown}{AC'} + \overset{\frown}{C'B} + \overset{\frown}{BA'} + \overset{\frown}{A'C} + \overset{\frown}{CB'} + \overset{\frown}{B'A}\right\}\\ &= \frac{1}{4}\cdot 360^\circ = 90^\circ.\end{aligned}$$

Hence AA' is an altitude in $\triangle A'B'C'$. Analogously, BB' and CC' are simultaneously bisectors in $\triangle ABC$ and altitudes in $\triangle A'B'C'$. This proves (a). Denote this common center by M.

(**b**) Since $\overset{\frown}{CB'} = \overset{\frown}{B'A}$ we have $\angle CC'B' = \angle B'C'A$. Because also $\angle C'PM = \angle C'PA (= 90^\circ)$ the triangles $C'PM$ and $C'PA$ are congruent. Hence $MP = PA = \frac{1}{2}MA$. In the same way $MQ = \frac{1}{2}MB$ and $MR = \frac{1}{2}MC$ and the assertion is immediate.

Solved also by S. Abramovich & T. Fujii & J. N. Wilson, E. Alkan (student, Turkey), J. Anglesio (France), R. Barbara (Lebanon), S. F. Barger, F. Bellot Rosado (Spain), G. Bennett, W. Blumberg, R. J. Chapman (U. K.), A. Coffman, H. W. Guggenheimer, J. Guichelaar (The Netherlands), L.-s. Hahn, J. G. Heuver (Canada), R. Holzsager, H. Kappus (Switzerland), I. Kastanas, D. C. Kay, K. S. Kedlaya (student), N. Komanda, J. H. Lee (student, Korea), H. M. Marston, A. Nijenhuis, G. Perz (Austria), C. G. Petalas (Greece), D. Schepler (student), R. A. Simon (Chile), A. Sinefakopoulos (student, Greece), R. Tauraso (Italy), R. S. Tiberio, M. Vowe (Switzerland), A. N. 't Woord (The Netherlands), R. L. Young, S. Yu & J. Yu, Anchorage Math Solutions Group, the MMRS group of Oklahoma State University, and the proposer.

Modular Exponentiation Hits a Rut

10324 [1993, 688]. *Proposed by William P. Wardlaw, United States Naval Academy, Annapolis, MD.*

Let a and m be positive integers and define the sequence $\langle x_n \rangle$ by $x_0 = 1$ and $x_{n+1} = a^{x_n}$. Show that there is a positive integer N such that $x_h \equiv x_k \pmod{m}$ whenever $N \leq h \leq k$.

Solution by Jerrold W. Grossman, Oakland University, Rochester, MI. We show that the sequence $1, a, a^a, a^{(a^a)}, \ldots$ is eventually constant modulo m. We proceed by induction on m. If $m = 1$, there is nothing to prove. Assume that the statement is true for all values less than m. If m is not a prime power, then by the inductive hypothesis and the Chinese Remainder Theorem, the sequence is eventually constant modulo each of the maximal prime-power divisors of m and therefore is constant modulo m. If m is a prime power and the underlying prime divides a, then the sequence is eventually 0 modulo m. Otherwise, a is relatively prime to m. By Euler's Theorem, the sequence is eventually constant modulo m because the induction hypothesis implies that it is eventually constant modulo $\phi(m)$, where ϕ is the Euler phi-function.

Editorial comment. Several solvers gave references to the literature where this result and generalizations can be found:

G. R. Blakley and I. Borosh, "Modular arithmetic of iterated powers", *Comput. Math. Appl.* 9 (1983), 567–581.

R. J. MacG. Dawson, "Towers of powers modulo m", *The College Mathematics Journal* 25 (1994), 22–28.

Torleiv Kløve, "On exponential recurring sequences", *Math. Scand.* 34 (1974), 44–50.

Daniel B. Shapiro and S. David Shapiro, "Iterated exponents in number theory", unpublished manuscript, 1986. (Available from D. B. Shapiro, Ohio State.)

Moreover, D. B. Shapiro noted that a similar result appeared as problem A-4 of the 1985 Putnam Examination and H. K. Krishnapriyan noted that the case $a = 2$ appeared in the 20th USA Mathematical Olympiad (see *Mathematics Magazine* 65 (1992), 205.)

Solved also by E. Alkan (student, Turkey), R. Barbara (Lebanon), K. Brown, P. Budney, R. J. Chapman (U. K.), G. Ehrlich, R. Holzsager, N. Jensen (Germany), T. Kløve (Norway), D. W. Koster, H. K. Krishnapriyan, J. H. Lindsey II, W. W. Meyer, A. Pedersen (Denmark), T. Ransford (Canada), D. B. Shapiro, T. White, A. N. 't Woord (The Netherlands), Anchorage Math Solutions Group, and the proposer.

A Sufficient Condition for Nilpotence

10339 [1993, 873]. *Proposed by Moshe Rosenfeld, Pacific Lutheran University, Tacoma, WA.*

Let A and B be complex matrices with $AB^2 - B^2A = B$. Prove that B is nilpotent.

Composite solution I by Joseph E. Higgins, Cadence Design Systems, Inc., San José, CA., and A. N. 't Woord, University of Technology, Eindhoven, The Netherlands. We prove the result for any algebra having a norm satisfying the multiplicative inequality $||XY|| \leq ||X||\ ||Y||$. Using induction, we first show that $2^{k-1}B^{2^k-1} = AB^{2^k} - B^{2^k}A$ for $k \in \mathbb{N}$. For $k = 1$ this is the given condition. If the equation holds for k, then multiplying on the left and right by B^{2^k} yields

$$2^{k-1}B^{2^{k+1}-1} = B^{2^k}AB^{2^k} - B^{2^{k+1}}A$$

$$2^{k-1}B^{2^{k+1}-1} = AB^{2^{k+1}} - B^{2^k}AB^{2^k},$$

which sum to the desired equation for $k + 1$.

By applying the multiplicative norm inequality to the resulting equation, we have for each k that

$$2^{k-1}||B^{2^k-1}|| \le 2||A||\,||B^{2^k-1}||\,||B||.$$

If B is not nilpotent, then $||B^{2^k-1}|| > 0$ for all k. Cancellation then yields $2^{k-2} \le ||A||\,||B||$ for all k, which is impossible. Hence B is nilpotent.

Solution II, including a characterization of B, by William C. Waterhouse, Pennsylvania State University, University Park, PA. Let B be a complex matrix. We prove that there exists A with $AB^2 - B^2A = B$ if and only if B is nilpotent and each Jordan block has odd size.

First, change basis to put B in Jordan form, with Jordan blocks $J_1, \dots, J_m$ along the diagonal. For each matrix A, let $A_1, \dots, A_m$ be the corresponding diagonal blocks. If the matrix A satisfies the required equation with B, then A_i satisfies it with J_i, since block multiplication yields $A_iJ_i^2 - J_i^2A_i$ as the ith diagonal block of $AB^2 - B^2A$. Conversely, given the A_i, the block diagonal matrix with the A_i on the diagonal and 0 elsewhere is a suitable choice for A.

Suppose some such A_i exists for J_i, with λ_i being the eigenvalue of J_i. Since size$(J_i)\lambda_i$ = trace(J_i) = trace$(A_iJ_i^2)$ − trace$(J_i^2A_i) = 0$, we conclude that $\lambda_i = 0$. Thus J_i is nilpotent, of index equal to its size. Induction from the equation yields $A_iJ_i^{2n} - J_i^{2n}A_i = nJ_i^{2n-1}$. Hence $J_i^{2n} = 0$ implies $J_i^{2n-1} = 0$, and J_i has odd size.

Conversely, suppose J is a nilpotent Jordan block of odd size. This yields a partial basis $e_0, \dots, e_{2n}$ with $Je_0 = 0$ and $Je_r = e_{r-1}$ for $r > 0$. Define $Ae_{2k} = (n-k)e_{2k+1}$ (with $Ae_{2n} = 0$) and $Ae_{2k-1} = -ke_{2k}$. It is easy to check that $AJ^2 - J^2A = J$.

This proof remains valid over an arbitrary field of characteristic zero. Although a field extension is needed to guarantee the existence of the Jordan canonical form, the fact that B turns out to be nilpotent shows that the Jordan form can be found over the original field.

Composite solution III, characterizing A, by David Callan, University of Wisconsin, Madison, WI, and the editors. Given B, let A_0 be the block diagonal matrix constructed in Solution II. The other choices for A are precisely the matrices $A_0 + Z$ with Z in the centralizer of B^2. It remains to find this centralizer. The structure of this set of matrices is most easily described in terms of the Jordan canonical form of B^2, which we assume to be written with Jordan blocks $J_1, \dots, J_m$, arranged in decreasing size along the diagonal. Let n_j be the number of rows in J_j. The matrix Z is then built from blocks Z_{ij} with $Z_{ij}J_j = J_iZ_{ij}$. It is easily seen that this requires that each Z_{ij} is characterized by the three properties: (i) the first column is zero except possibly for the first entry; (ii) the last row is zero except possibly for the last entry; (iii) the (k, l)-entry is equal to the $(k+1, l+1)$-entry so that the matrix entries are *constant on diagonals.* Thus, there are $\min(n_i, n_j) = n_{\max(i,j)}$ parameters in the description of Z_{ij}, and hence the dimension of the centralizer is

$$\sum_{j=1}^{m}(2j-1)n_j.$$

Editorial comment. S. C. Locke also characterized the possible choices of A for a given B as in Solution III. Gertrude Ehrlich, Charles Lanski, and Harvey Schmidt, Jr., independently, also gave a characterization of B as in solution II. The National Security Agency Problems Group cited a more general lemma of Jacobson for commutators of complex matrices: with $[X, Y]$ denoting $XY - YX$, the condition $[Y, [X, Y]] = 0$ implies that $[X, Y]$ is nilpotent (see H. Flanders, "Methods of proof in linear algebra", this MONTHLY 63(1956), 1–15, especially section 12, p. 10). The same result was mentioned by Saieed Akbari with a reference to I. N. Herstein, *Topics in Algebra*, Wiley, 1975, Lemma 6.84, p. 316. Thus, as noted by several solvers, B^2 could be replaced by an arbitrary polynomial in B. Robin

Harte cited I. Singer and J. Wermer, "Derivation on commutative normed algebras", *Math. Ann.* 129(1955), 260–264 as the source for the more general result: if $[Y, [X, Y]] = 0$ in a Banach algebra, then $[X, Y]$ is quasi-nilpotent, which in the finite-dimensional case means nilpotent. A.A. Jagers provided a counterexample over $\mathbb{Z}_3$:

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Solved by 44 readers, including those cited, and the proposer. One incorrect solution was received.

Pythagorean Orthogonality

10340 [1993, 873]. *Proposed by Richard Bagby, New Mexico State University, Las Cruces, NM.*

For a normed linear space $\mathbf{X}$ and $x \in \mathbf{X}$, define

$$P(x) = \left\{ y \in \mathbf{X} : \|x + y\|^2 = \|x\|^2 + \|y\|^2 \right\}.$$

If the norm in $\mathbf{X}$ comes from an inner product, then each $P(x)$ is invariant under multiplication by real numbers. Is the converse true?

Solution I by Charles R. Diminnie, Saint Bonaventure University, Saint Bonaventure, NY. The converse is true. Indeed, it can be found in theorems 5.1 and 5.2 of R. C. James, "Orthogonality in normed linear spaces", *Duke Math. J.* 12 (1945), 291–302. In that paper, the *Pythagorean orthogonality* of vectors x and y is defined by the relation $\|x + y\|^2 = \|x\|^2 + \|y\|^2$ and denoted $x \perp y$. This relation is said to be homogeneous if $x \perp y$ implies $x \perp ay$ for all $\alpha \in \mathbb{R}$. The cited theorems show that homogeneity of $x \perp y$ characterizes inner product spaces, which is the desired result. The proof is based on

Lemma. *For every $x, y \in \mathbf{X}$, there is a real number α such that $x \in P(\alpha x + y)$.*

Proof. If $x = 0$, the result is true since $0 \in P(z)$ for every $z \in \mathbf{X}$. Assume $x \neq 0$ and consider the function

$$\begin{aligned} f(\tau) &= \|(1 + \tau)x + y\|^2 - \|\tau x + y\|^2 - \|x\|^2 \\ &= \left(\|(1 + \tau)x + y\| - \|\tau x + y\| \right)\left(\|(1 + \tau)x + y\| + \|\tau x + y\| \right) - \|x\|^2 \end{aligned}$$

For $\tau > 0$,

$$\left\|(1 + \tau)x + y\right\| - \left\|\tau x + y\right\| = \left\|x + (1 + \tau)^{-1}y\right\| + \left\|\tau x + \tau(1 + \tau)^{-1}y\right\| - \left\|\tau x + y\right\|$$

with

$$\left| \left\|\tau x + \tau(1 + \tau)^{-1}y\right\| - \left\|\tau x + y\right\| \right| \leq \left\|(1 + \tau)^{-1}y\right\|.$$

Also, for $\tau < -1$,

$$\left\|(1 + \tau)x + y\right\| - \left\|\tau x + y\right\| = -\left\|x + (1 + \tau)^{-1}y\right\| + \left\|\tau x + \tau(1 + \tau)^{-1}y\right\| - \left\|\tau x + y\right\|$$

with

$$\left| \left\|\tau x + \tau(1 + \tau)^{-1}y\right\| - \left\|\tau x + y\right\| \right| \leq \left\|(1 + \tau)^{-1}y\right\|.$$

It follows that $f(\tau) \to \pm\infty$ as $\tau \to \pm\infty$. Since $f(\tau)$ is continuous on $\mathbb{R}$, there is a real number α such that $f(\alpha) = 0$.

Now assume that $P(z)$ is invariant under real multiplication for every $z \in \mathbf{X}$. Given any x and y, the lemma implies that there is a real number α such that $x \in P(\alpha x + y)$. By

invariance under real multiplication, it follows that $-\alpha x$, $(1-\alpha)x$, and $-(1+\alpha)x$ are also in $P(\alpha x + y)$. Then

$$\|y\|^2 = \alpha^2 \|x\|^2 + \|\alpha x + y\|^2 ,$$
$$\|x+y\|^2 = (1-\alpha)^2 \|x\|^2 + \|\alpha x + y\|^2 ,$$

and

$$\|x-y\|^2 = (1+\alpha)^2 \|x\|^2 + \|\alpha x + y\|^2 .$$

Therefore,

$$\|x+y\|^2 + \|x-y\|^2 = 2\big(\|x\|^2 + \|y\|^2 \big)$$

and the conclusion follows since this parallelogram law is known to characterize inner products (see P. Jordan and J. von Neumann, "On inner products in linear metric spaces", *Ann. of Math.* 36 (1935), 719–723).

Solution II by Lawrence Crone and Richard Holzsager, The American University, Washington, DC. Set $\langle u, v\rangle = \big(\|u+y\|^2 - \|u\|^2 - \|v\|^2 \big)/2$. If the norm comes from an inner product, this must be it. If we can show that $\langle u, v\rangle$ is bilinear, we will be done, since symmetry and positive definiteness are clear. First, let's see that the result holds for two-dimensional spaces (dimension one is trivial). Let u and w be independent vectors. If $v = \cos\theta u + \sin\theta w$, then $\langle u, v\rangle > 0$ when $\theta = 0$ and $\langle u, v\rangle < 0$ when $\theta = \pi$, so for some θ, $\langle u, v\rangle = 0$. Fix this vector v; then u and v satisfy $\|u+v\|^2 = \|u\|^2 + \|v\|^2$. By the hypothesis, $\|\alpha u + \beta v\|^2 = \alpha^2\|u\|^2 + \beta^2\|v\|^2$, for all α and β. Then for and λ, μ, ρ, and σ,

$$\begin{aligned} 2\langle \lambda u + \mu v, \rho u + \sigma v\rangle &= \|(\lambda+\rho)u + (\mu+\sigma)v\|^2 - \|\lambda u + \mu v\|^2 + \|\rho u + \sigma v\|^2 \\ &= \big((\lambda+\rho)^2 - \lambda^2 - \rho^2 \big) \|u\|^2 + \big((\mu+\sigma)^2 - \mu^2 - \rho^2 \big) \|v\|^2 \\ &= 2\big(\lambda\rho \,\|u\|^2 + \mu\sigma \,\|v\|^2 \big) \end{aligned}$$

This clearly defines an inner product, so at least on two-dimensional subspaces, the norm comes from an inner product.

If we knew the result for three dimensional spaces, we would be done since only three vectors appear in verifying linearity of $\langle \_, \_\rangle$ in either variable. For the remainder of the proof, $\mathbf{X}$ will be assumed to be three dimensional. Introduce $S = \{u : \|u\| = 1\}$ with respect to the given norm on $\mathbf{X}$. Choose a coordinate system on $\mathbf{X}$ and construct the standard Euclidean norm with respect to this coordinate system. The remainder of the proof will be expressed in terms of this Euclidean geometry. Let $x \in S$ be a point whose Euclidean norm is largest, and rotate coordinates so that the X-axis passes through x. Since the desired result has been proved for two dimensional subspace, any plane through the X-axis meets S in an ellipse. Furthermore, x is one end of the major axis of this ellipse (if the intersection is a circle, nothing prevents us from calling any diameter the major axis). These ellipses are then completely determined by the intersection of S with the plane perpendicular to the X-axis. Again, by the two dimensional result, this curve is an ellipse, forcing S to be an ellipsoid. Indeed, by choosing the Y axis to pass through a point on the curve of maximal Euclidean norm, the equation of S will be in the usual diagonal form. This is sufficient to express $\langle \_, \_\rangle$ as an inner product.

Editorial comment. Mahmood A. Al-Abbas cited Theorem 5.2 of Mahlon M. Day, "Some characterizations of inner-product spaces", *Trans. Amer. Math. Soc.* 62 (1947), 320–337, which says that it suffices that each $P(x)$ be invariant under multiplication by -1 in order to show that $\mathbf{X}$ is an inner-product space. For simplicity of exposition, solutions have been given for normed spaces over $\mathbb{R}$. The cited literature studied these properties also for normed spaces over $\mathbb{C}$, allowing the stated result to be proved in that interpretation also.

Solved also by M. A. Al-Abbas (Bahrain), D. Muraz & G. Muraz (France), A. Pedersen (Denmark), A. Tissier (France), A. N. 't Woord (The Netherlands), and the proposer.

REVIVALS

An Insufficient Condition for Primality

10268 [1992, 958; 1995, 557]. *Proposed by Ondrej Šuch (student), Queens University, Kingston, Ontario, Canada.*

Define a sequence $\langle a_n \rangle$ for $n \in \mathbb{N}$ by

$$a_0 = 3 \qquad a_1 = 0 \qquad a_2 = 2$$
$$a_{n+3} = a_{n+1} + a_n \quad (n \in \mathbb{N}).$$

If p is a prime, show that $p | a_p$.

Editorial comment. Naoki Sato noted that the identity $a_n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}a_{3k}$ in the selected solution is incorrect. The correct consequence of that argument is $a_n = \sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}a_{3k}$, but this gives no new information when n is an odd prime.

Since several correct solutions were outlined in the *Editorial Comment*, it is not necessary to give another solution here. We can, however, update the reference to the problem of Paul S. Bruckman mentioned there. It is *Fibonacci Quarterly* H-477 [1993, 188; 1994, 474]. The published solution to that problem by H.-J. Seiffert includes a solution of this problem.

A Cubic Relative of the AGM

10281 [1993, 76; 1996, 181]. *Proposed by Jonathan M. Borwein, University of Waterloo, Waterloo, Ontario, Canada.*

For $a > 0$ and $b > 0$, let

$$I(a, b) = \int_0^\infty \frac{t\,dt}{\sqrt[3]{(a^3 + t^3)(b^3 + t^3)^2}}.$$

(**a**) Show that

$$I(a, b) = I\left(\frac{a + 2b}{3}, \sqrt[3]{b\frac{a^2 + ab + b^2}{3}}\right).$$

(**b**) Show that the iteration which has $a_0 = a$ and $b_0 = b$ and

$$a_{n+1} = \frac{a_n + 2b_n}{3}$$
$$b_{n+1} = \sqrt[3]{b_n\frac{b_n^2 + a_nb_n + a_n^2}{3}}$$

converges to $I(1, 1)/I(a, b)$.

Editorial comment. The paper, John A. Macdonald, Jonathan M. Borwein, and Petr Lisoněk, "Solution to Problem AMM 10281", *MapleTech*, to appear, gives a solution of this problem that requires only changes of variable that can be easily checked by a computer

mathematics system. The manuscript is expected to be available from the home page of the Centre for Experimental and Constructive Mathematics (the current affiliation of the proposer) on the World Wide Web (**http://www.cecm.sfu.ca**).

The key lemma establishes the formula

$$I\left(1, \sqrt[3]{1-\gamma^3}\right) = \int_1^\infty \frac{dx}{\sqrt{(x-1)\left((x+3)x^2 - 4\gamma^3\right)}} \tag{$*$}$$

for any $\gamma < 1$. Several substitutions and a splitting of the original interval of integration into two parts are used in obtaining the formula $(*)$ from the original definition of $I(a, b)$. The desired change of γ can be verified by another change of variable in this integral.

A Recurrence Related to Counting Involutions

10347 [1993, 951; 1995, 844]. *Proposed by T. S. Nanjundiah, University of Mysore, Mysore, India.*

For integer $n \geq 1$, define real numbers R_n by

$$R_1 = 1 \qquad R_{k+1} = 1 + \frac{k}{R_k} \quad (k \geq 1).$$

Prove that

$$\sqrt{n - \frac{3}{4}} + \frac{1}{2} \leq R_n \leq \sqrt{n + \frac{1}{4}} + \frac{1}{2}$$

for $n \geq 1$.

Editorial comment. Tim Keller observed that the generalization to arbitrary R_1 mentioned in the *Editorial comment* could not possibly be correct. We repeat it here with equations labeled and the missing term in $(**)$ inserted.

With arbitrary $R_1 \geq 1$ and subsequent R_k defined by

$$R_{k+1} = R_1 + \frac{k}{R_k} \tag{$*$}$$

one has

$$\sqrt{n - 1 + \frac{R_1^2}{4}} + \frac{R_1}{2} \leq R_n \leq \sqrt{n + \frac{R_1^2}{4}} + \frac{R_1}{2} \tag{$**$}$$

for $n \geq 1$.

Note that R_1 has also replaced 1 in the recurrence $(*)$. Keller noted that, had the original recurrence been used and only the starting value changed, then the original inequality would be satisfied for all large n — except for one special negative value of R_1 — and this would contradict most variants of $(**)$. His proof resembles the treatment in Leo Moser & Max Wyman, "On solutions of $x^d = 1$ in symmetric groups", *Canadian J. Math.* 7 (1955), 159–168 that had been mentioned in the original solution.

REVIEWS

Edited by **Darrell Haile**
Indiana University, Bloomington, IN 47405

Algebraic Topology: A First Course, by William Fulton, Springer-Verlag, 1995, (Graduate Texts in Mathematics no. 153), xvii + 430, $39.50

Reviewed by **William S. Massey**

This textbook grew out of undergraduate courses the author taught at Brown University and the University of Chicago. The author states that the book is designed for students of mathematics or science who are not aiming to become practicing algebraic topologists, without, however, discouraging those who are budding topologists.

There are already many textbooks for a first course in algebraic topology. It may be a slight exaggeration to say that there is a consensus about the topics that should be taken up in such a course. However, the majority of the current texts seem to be built around the following topics: the fundamental group, covering spaces, the classification of compact 2-manifolds, singular homology and cohomology theory, the use of simplicial or CW-complexes to compute homology groups, and various applications of the techniques developed to actual problems. Of course many of these texts also treat a few other topics, but usually those listed form the core of the book. Given this quasi-consensus, it is natural to fear that any new textbooks on the subject will be only a slight variation on the existing texts. Fortunately, that is not the case for the book under review: Fulton has written a book that *is* different. We will try to make clear how this book is different as we discuss the topics taken up.

Following the standard practice in algebraic topology textbooks on this level, the main emphasis is on homology and cohomology theory. Throughout this book the only homology groups considered are singular homology groups with integer coefficients, defined by using cubes rather than simplexes. The only cohomology groups considered are the de Rham cohomology groups of a differentiable manifold, defined using C^∞ differential forms. There are a few minor exceptions to these statements that we will note later on. The relative homology and cohomology groups of a pair (X, A) consisting of a topological space X and a subspace A are not considered. As a result, the exact homology or cohomology sequence of such a pair is never used. Instead there is much use of the Mayer-Vietoris exact sequence for a space or manifold that is the union of two open subsets. The homomorphism induced on homology groups by a continuous map is defined, and is proved to be invariant under homotopies. The homomorphism induced on cohomology groups by a differentiable map of one manifold to another is also defined.

The reader is introduced to these ideas about homology and cohomology very gently and gradually. First, the homology and cohomology groups in dimensions 0

and 1 are defined for an open subset of the plane and applied to various problems. Then it is pointed out that the definitions of the 0- and 1-dimensional homology groups apply without change to an arbitrary topological space. After the introduction of differentiable surfaces (i.e., 2-dimensional manifolds), 2-dimensional cohomology groups are defined (but not 2-dimensional homology groups). Finally, in the last part of the book, the homology groups in all dimensions are introduced for arbitrary topological spaces, and cohomology groups in all dimensions are defined for arbitrary differentiable manifolds.

The first ten chapters of the book (150 pages) are concerned mainly with the topology of subsets of the plane. The book commences with a discussion of path integrals in the plane, essentially a review of a topic in advanced calculus. This is followed by the definition of the winding number of a closed path about a point. Several applications of winding numbers are given: a proof of the fundamental theorem of algebra, the Brouwer fixed point theorem for a disc, the Borsuk-Ulam theorem for antipode preserving self maps of the circle and the ham sandwich theorem. Next comes the definition of the 0- and 1-dimensional de Rham cohomology groups and the construction of part of the Mayer-Vietoris exact sequence for two open subsets of the plane. This machinery is sufficient to prove the Jordan Curve theorem in full generality; Brouwer's theorem on "Invariance of Domain" for open subsets of the plane is an easy corollary. These theorems should give the student some idea of the power of these ideas. Other topics in these chapters are the definitions of the 0- and 1-dimensional homology groups and the construction of the Mayer-Vietoris exact homology sequence in dimensions 0 and 1, the definition of the index of a singularity of a vector field in the plane or on a surface, and various classical theorems about such indices of singularities.

The middle part of the book has eight chapters that are devoted mainly to the fundamental group, covering spaces, and the topology of surfaces. The discussion of the fundamental group and the theory of covering spaces is standard, as is the proof of the classification theorem for compact, orientable surfaces (the case of non-orientable surfaces is relegated to an exercise). The fundamental group and first homology group of any surface are determined. Also included is a discussion of the de Rham cohomology groups $H^0(X)$, $H^1(X)$, and $H^2(X)$ of a (differentiable) surface X, the "cup product" pairing $H^1(X) \times H^1(X) \to H^2(X)$ defined by the product of differential forms, and the Mayer-Vietoris exact cohomology sequence for two open subsets of such a surface.

The next part of the book is perhaps its most unusual feature: it consists of three chapters (54 pages) devoted to Riemann surfaces. After the necessary basic definitions, lemmas, etc., these chapters culminate in the statement and proof of the famous Riemann-Roch theorem. Actually this theorem is proved only for Riemann surfaces that arise from a plane algebraic curve, thus avoiding the necessity of proving the existence of a non-constant meromorphic function on an arbitrary Riemann surface. There is also a proof of the Abel-Jacobi theorem, which relates the possible integrals on a Riemann surface to its topology. Usually the Riemann-Roch theorem comes up only in a specialized graduate course on algebraic curves or Riemann surfaces. Most universities rarely offer such courses.

The last section, entitled "Higher Dimensions," consists of three chapters (50 pages). Here the homology and cohomology groups are defined in all dimensions and the Mayer-Vietoris exact sequences for homology and cohomology are constructed in full generality. For an arbitrary differentiable manifold M^n the classical de Rham theorem is proved in the following form: integration of q-forms over

differentiable singular q-chains defines a natural homomorphism

$$H^q(M^n) \to Hom[H_q(M^n), \mathbf{R}],$$

which is an isomorphism. From the author's point of view, this famous theorem is the basic relation between homology and cohomology. The degree of a continuous map of an n-sphere to itself is defined, and various standard theorems are proved as a consequence. In another section, the Jordan-Brouwer separation theorem about an $(n - 1)$-dimensional sphere in $\mathbf{R}^n$ is proved, along with some corollaries.

In the last chapter the author introduces de Rham cohomology with compact supports by using differential forms with compact supports. The q-dimensional cohomology group of M with compact supports is denoted by $H_c^q(M)$. One may then define a multiplication or pairing $H^k(M) \times H_c^{n-k}(M) \to \mathbf{R}$ for an orientable n-manifold M as follows: take the product of a representative closed k-form and a representative closed $(n - k)$ form with compact support. Integrate this closed n-form over the entire manifold; the result is a real number, which is independent of the choices of representatives. This multiplication defines a homomorphism

$$H^k(M) \to Hom[H_c^{n-k}(M), \mathbf{R}].$$

It is now proved that this homomorphism is an isomorphism for any n-dimensional orientable differentiable manifold. This result is a special case of the general Poincaré duality theorem. Tacked on at the end of this last chapter is a section on simplicial complexes. It is proved that the usual homology groups of a finite simplicial complex are isomorphic to the singular homology groups of the underlying topological space. This standard result is not used anywhere in the book.

There are several appendices at the end of the book. The last of these appendices is devoted to a proof of the general Borsuk-Ulam theorem about antipodal maps of a sphere to a sphere. This proof uses homology groups with the integers mod 2 as coefficients. It is the only place in the book that homology groups with other than integer coefficients are used.

This book is noteworthy because a lot of motivational material is included to help the student understand the various concepts that are introduced. Probably most textbooks on algebraic topology do not have enough motivational material. To the student, some of these books must seem to consist mostly of algebraic formalisms. Fulton has chosen to develop the subject by means of examples and applications that were historically important, although of course he has not tried to recapitulate the historical development of the subject.

On the other hand, some teachers who are about to give a course on algebraic topology may decide that the book moves too slowly and does not get to some of the important ideas of algebraic topology quickly enough, or perhaps does not get to some of them at all. Unfortunately, students nowadays are under great pressure to assimilate a lot of material in a short time, and can not always afford the luxury of a leisurely treatment of a subject, with all the examples and applications it would be desirable to include. This applies not only to graduate students who are struggling to get a Ph.D. in mathematics; for example, many theoretical physicists nowadays feel it is necessary to learn some of the more subtle ideas of algebraic topology, but have only limited time available.

In this book, as in most books on this subject, many of the motivational examples involve low dimensional topological spaces (i.e., 1, 2, or 3 dimensional). This is in accord with the historical development of algebraic topology. Most of the definitions and techniques of homology and cohomology theory, which apply to all

dimensions, are based on low dimensional examples where geometric intuition is able to guide us. In fact, one of the main tasks of algebraic topology is to deal rigorously with various problems in higher dimensions that are often more or less trivial in low dimensions. An amazing result of this exploration of higher dimensional topology is that often totally new and unexpected phenomena are discovered that have no analogue in lower dimensions. For example, for most values of $n > 6$, the n-dimensional sphere admits several different (i.e., non-isomorphic) structures as a differentiable manifold, while for $n = 1$, 2, or 3 there is only one possibility for a differentiable manifold structure on the n-dimensional sphere (up to isomorphism). The case $n = 4$ is still unsolved. Unfortunately, it is not possible to explain exciting examples like this in a beginning text, such as the one under review. Too much machinery is required.

Fulton has done genuine service for the mathematical community by writing a text on algebraic topology that is genuinely different from the existing texts. Each time a text such as this is published we more truly have a real choice when we pick a book for a course or for self-study. The author, who is an expert in algebraic geometry, has given us his own personal idiosyncratic vision of how the subject should be developed.

Department of Mathematics
Yale University
New Haven, CT 06520

Mathematical rigor is like clothing; in its style it ought to suit the occasion, and it diminishes comfort and restrains freedom of movement if it is either too loose or too tight.

Mathematical Intelligencer, Vol. 13, No. 1, Winter 1991

I'm sorry to say that the subject I most disliked was mathematics. I have thought about it. I think the reason was that mathematics leaves no room for argument. If you made a mistake, that was all there was to it.

Malcolm X

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1-4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

General. *Thought and Knowledge: An Introduction to Critical Thinking, Third Edition.* Diane F. Halpern. Lawrence Erlbaum Associates, 1996, $37.50 (P), set; xiv + 430 pp, $29.95 (P), [ISBN 0-8058-1494-9; 0-8058-1493-0]; *Thinking Critically About Critical Thinking*, ix + 205 pp, $19.95 (P). [ISBN 0-8058-2032-9] Introduction to the psychological processes of thinking, reasoning, and decision making. Workbook contains exercises (such as logic puzzles). DH

Reference, P, L. *Mathematics Handbook for Science and Engineering.* Lennart Råde, Bertil Westergren. Birkhäuser Boston, 1995, 539 pp, $45. [ISBN 0-8176-3858-X] Definitions, theorems, formulas, graphs, and tables. Material from discrete mathematics, geometry, linear algebra, analysis, probability, statistics, optimization, and numerical analysis. AO

Finite Mathematics, T(13). *Fundamentals of Mathematics, Seventh Edition.* William M. Setek, Jr. Prentice Hall, 1996, xxiii + 744 pp. [ISBN 0-02-409280-0] New: a chapter on fundamentals of problem solving, collaborative learning projects, and revised exercises. (*Sixth Edition*, TR, November 1993.) DH

Foundations, T(14–15: 1), S*, L. *An Accompaniment to Higher Mathematics.* George R. Exner. Undergrad. Texts in Math. Springer-Verlag, 1996, xvii + 198 pp, $29.95. [ISBN 0-387-94617-9] Innovative text for "transition" course or supplement for first proof course. First three chapters cover active reading, construction of examples, informal and formal language and proof; numerous exercises. Final chapter provides additional practice in form of "laboratories" on sets and functions. KES

Discrete Mathematics, P. *Threshold Graphs and Related Topics.* N.V.R. Mahadev, U.N. Peled. Annals of Disc. Math., V. 56. North-Holland (US Distr: Elsevier Science), 1995, xiii + 543 pp, $156.25. [ISBN 0-444-89287-7] Self-contained book on threshold graphs. Includes open problems and research ideas. LC

Calculus, S(13). *Discovering Calculus with Mathematica.* Cecilia A. Knoll, *et al.* Wiley, 1995, viii + 340 pp, $20.95 (P). [ISBN 0-471-00976-8] Supplement for traditional calculus course. Uses *Mathematica*'s graphical, symbolic, and numerical capabilities to solve standard calculus problems. DH

Dynamical Systems, T(16–17: 2), P. *An Introduction to Symbolic Dynamics and Coding.* Douglas Lind, Brian Marcus. Cambridge Univ Pr, 1995, xvi + 495 pp, $27.95 (P); $54.95. [ISBN 0-521-55900-6; 0-521-55124-2] Symbolic dynamics is a part of dynamical systems, and arose from studying the latter by discretizing space as well as time. This book offers an introduction to the main ideas, techniques, and some applications. LC

Numerical Analysis, P*, L.** *Real Computing Made Real: Preventing Errors in Scientific and Engineering Calculations.* Forman S. Acton. Princeton Univ Pr, 1996, xv + 259 pp, $29.95. [ISBN 0-691-03663-2] A witty, useful book on the practicalities of scientific computing. Provides a wealth of hints and advice on avoiding numerical pitfalls (e.g., loss of significant digits, iterative instabilities, convergence

to extraneous roots). Many examples and exercises (with answers). AO

Numerical Analysis, P. *Two Dimensional Spline Interpolation Algorithms.* Helmuth Späth. AK Peters, 1995, viii + 304 pp, $59.95. [ISBN 1-56881-017-2] A practical introduction to two-dimensional spline interpolation that includes many FORTRAN–77 subroutines. Similar in spirit and approach to the author's book *One Dimensional Spline Interpolation Algorithms* (TR, March 1996). AO

Numerical Analysis, P*, L*. *Accuracy and Stability of Numerical Algorithms.* Nicholas J. Higham. SIAM, 1996, xxviii + 688 pp, $39 (P). [ISBN 0-89871-355-2] In-depth treatment combining derivations of algorithms, perturbation theory, and rounding error analysis for many standard matrix computation algorithms. Other topics: floating point summation, block LU factorization, condition number estimation, the Sylvester equation, powers of matrices, finite precision behavior of stationary iterative methods, Vandermonde systems, and fast matrix multiplication. Extensive citations to the research literature. AO

Numerical Analysis, T(17: 1). *Numerical Partial Differential Equations: Finite Difference Methods.* J.W. Thomas. Texts in Appl. Math., V. 22. Springer-Verlag, 1995, xx + 437 pp, $49.50. [ISBN 0-387-97999-9] Finite difference methods for parabolic and hyperbolic PDEs. Emphasizes both theoretical and practical aspects. Stresses use of technology (e.g., symbolic computing, graphics). AO

Operator Theory, P. *Methods of Noncommutative Analysis: Theory and Applications.* Vladimir E. Nazaikinskii, Victor E. Shatalov, Boris Yu. Sternin. Stud. in Math., V. 22. Walter de Gruyter, 1996, x + 373 pp, DM 198. [ISBN 3-11-014632-0] Introduction to the calculus of functions of noncommuting operators. Applications include electromagnetic waves in plasma and geostrophic wind equations. JO

Analysis, T(17: 1), L*. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics.* Nico M. Temme. Wiley, 1996, xii + 374 pp, $54.95. [ISBN 0-471-11313-1] A nice introduction to many of the special functions that arise in physics, engineering, and probability theory. AO

Geometry, T(16-17: 1), P, L*. *Convexity.* Roger Webster. Oxford Univ Pr, 1994, xvii + 444 pp, $69.96. [ISBN 0-19-853147-8] An attractive and appealing introduction to the theory and applications of convex sets and functions. Motivates concepts with geometric examples. Requires only an elementary background in linear algebra and analysis. Provides solutions to almost all of the exercises. AO

Geometry, P, L. *Geometric Tomography.* Richard J. Gardner. Ency. of Math. & Its Applic., V. 58. Cambridge Univ Pr, 1995, xvi + 424 pp, $64.95. [ISBN 0-521-45126-4] Deals with the retrieval of information about a geometric object from data about its planar cross-sections, its projections, or both. Includes brief introductions, accessible to advanced undergraduates, to the basic concepts, and a list of 72 unsolved problems. AO

Optimization, T(17: 2), L. *Nonlinear Programming.* Dimitri P. Bertsekas. Athena Scientific, 1995, x + 646 pp, $69. [ISBN 1-886529-14-0] An up-to-date and mathematically rigorous introduction to nonlinear programming. Covers both traditional topics (e.g., Lagrange multiplier theory, descent methods) and recent developments (e.g., interior point methods, neural network training). Presents both variational and geometric approaches to duality theory. AO

Optimization, P. *Global Optimization in Action.* János D. Pintér. Nonconvex Optim. & Its Applic., V. 6. Kluwer Academic, 1996, xvii + 478 pp, $198. [ISBN 0-7923-3757-3] Presents adaptive partition strategies for solving Lipschitzian and continuous global optimization problems, and some applications of these techniques. AO

Optimization, T(17: 1), L. *Introduction to Global Optimization.* Reiner Horst, Panos M. Pardalos, Nguyen V. Thoai. Nonconvex Optim. & Its Applic., V. 3. Kluwer Academic, 1995, xii + 318 pp, $174. [ISBN 0-7923-3556-2] Introduces techniques for solving a variety of constrained global optimization problems including nonconvex quadratic programming, general concave minimization, network optimization, Lipschitz optimization, and DC programming. Assumes knowledge of elementary real analysis, linear algebra, and classical linear programming. AO

Optimization, T(16: 1), L. *An Introduction to Optimization.* Edwin K.P. Chong, Stanislaw H. Zak. Ser. in Disc. Math. & Optim. Wiley, 1996, xiii + 409 pp, $59.95. [ISBN 0-471-08949-4] Techniques for unconstrained optimization, linear programming, and nonlinear constrained optimization. Includes some material on new approaches and applications (e.g., Karmarkar's algorithm, neural network training, genetic algorithms). First part reviews topics from multivariable calculus and linear algebra. AO

Optimal Control, T(17: 2), L. *Dynamic Programming and Optimal Control.* Dimitri P. Bertsekas. Athena Sci, 1995, $99.50 set, [ISBN 1-886529-11-6] set. *Volume I*, xi + 387 pp; *Volume II*, xi + 292 pp. For engineers, operations researchers, economists, and applied mathematicians. *Volume I* focuses on finite horizon problems and modeling; *Volume II* focuses on infinite horizon problems, mathematical analysis, and computation. AO

Elementary Statistics, T(13: 1). *Seeing Through Statistics.* Jessica M. Utts. Duxbury Pr, 1996, xvi + 464 pp, $39. [ISBN 0-534-25776-3] An easy to read introduction to statistics using real data. Calculations are shown step-by-step with formulas available at chapter's end. Intended for social scientists. Contains case studies and exercises. DH

Mathematical Computing, S, P, C. *The Practical Approach Utilities for Maple, Maple V, Release 3.* Darren Redfern. Springer-Verlag, 1995, viii + 312 pp, $69 (P), with disk. [ISBN 0-387-14221-5] Supplementary Maple commands for pattern matching, for manipulating lists, strings, arrays, and other data structures, for drawing and manipulating graphs, and for formatted I/O. The accompanying book describes these commands as well related built-in Maple commands. AO

Mathematical Computing, P. *Maple V: Programming Guide.* M.B. Monagan, *et al.* Springer-Verlag, 1996, xii + 379 pp, $34 (P). [ISBN 0-387-94537-7] A practical, introductory guide to writing Maple programs. Examples illustrate symbolic, numeric, and graphics applications. AO

Programming, S(13), L*. *The Little Schemer, Fourth Edition.* Daniel P. Friedman, Matthias Felleisen. MIT Pr, 1996, xiii + 196 pp, $16.50 (P). [ISBN 0-262-56099-2] A revised and updated edition of the deservedly popular book formerly known as *The Little LISPer* (TR, June–July 1988). Introduces programs as recursive functions. AO

Programming, S(13), L*. *The Seasoned Schemer.* Daniel P. Friedman, Matthias Felleisen. MIT Pr, 1996, xiii + 210 pp, $18.50 (P). [ISBN 0-262-56100-X] A sequel to *The Little Schemer.* Introduces additional concepts: functions as values, change of state, and exceptional cases. AO

Programming, T(15–16: 1). *Practical Parallel Programming.* Gregory V. Wilson. Sci. & Eng. Comput. MIT Pr, 1995, ix + 564 pp, $50. [ISBN 0-262-23186-7] Well-written introduction to general techniques for programming parallel processors. Appendix on history, annotated bibliography, glossary. JO

Programming, S(14), L. *C++: The Core Language.* Gregory Satir, Doug Brown. O'Reilly & Associates, 1995, xix + 207 pp, $19.95 (P). [ISBN 1-56592-116-X] For C programmers who wish to learn C++. Presents only a subset of the language, but sidebars provide overviews of features not covered. AO

Algorithms, T(17: 1), L. *Randomized Algorithms.* Rajeev Motwani, Prabhakar Raghavan. Cambridge Univ Pr, 1995, xiv + 476 pp, $39.95. [ISBN 0-521-47465-5] Basic concepts in the design and analysis of randomized algorithms. First part presents basic tools and techniques. Second part focuses on applications: data structures, graph algorithms, geometric algorithms, number theoretic algorithms, counting algorithms, parallel and distributed algorithms, and online algorithms. Assumes background in algorithm analysis and probability theory. AO

Computer Science, T(17: 1), P, L. *Pseudorandomness and Cryptographic Applications.* Michael Luby. Comp. Sci. Notes. Princeton University Pr, 1996, xvi + 234 pp, $24.95 (P). [ISBN 0-691-02546-0] Rigorous definitions and proofs related to cryptography. First half of book shows how to construct a pseudorandom number generator from any one-way function. The second half shows how to build other useful cryptographic primitives. AO

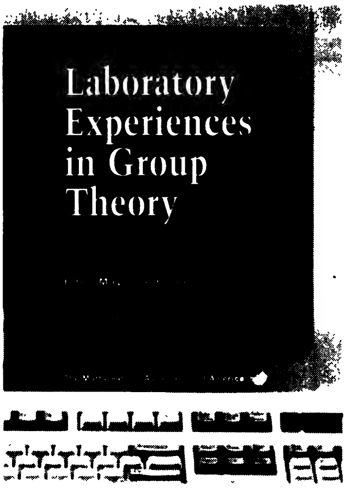
Applications (Economics), T(16: 1), L*. *The Mathematics of Financial Derivatives: A Student Introduction.* Paul Wilmott, Sam Howison, Jeff Dewynne. Cambridge Univ Pr, 1995, xiii + 317 pp, $24.95 (P); $49.95. [ISBN 0-521-49789-2; 0-521-49699-3] An introduction to the principles and practice of derivative pricing from an applied mathematician's perspective. Modeling, analysis, and numerical computation techniques based on PDEs. Knowledge of calculus and probability theory is required. AO

Applications, P. *Lagrange and Finsler Geometry: Applications to Physics and Biology.* Eds: P.L. Antonelli, R. Miron. Fund. Theories of Physics, V. 76. Kluwer Academic, 1996, ix + 279 pp, $137. [ISBN 0-7923-3873-1] 24 papers on applications of differential geometry.

Reviewers

LC: Laura Chihara, St. Olaf; DH: Deanna Haunsperger, Carleton; JO: Jeff Ondich, Carleton; AO: Arnold Ostebee, St. Olaf; KES: Kay E. Smith, St. Olaf.

THE MATHEMATICAL ASSOCIATION OF AMERICA



Laboratory Experiences in Group Theory

A Manual to be Used with *Exploring Small Groups*

Ellen Maycock Parker

Series: Classroom Resource Materials

A lab manual with software for introductory courses in group theory or abstract algebra

Laboratory Experiences in Group Theory is a workbook of 15 laboratories designed to be used with the software *Exploring Small Groups* as a supplement to the regular textbook in an introductory course in group theory or abstract algebra. Written in a step-by-step manner, the laboratories encourage students to discover the basic concepts of group theory and to make conjectures from examples that are easily generated by the software. The labs can be assigned as homework or can be used in a structured laboratory setting. Since the software is user-friendly and the laboratories are complete, students and faculty should have no difficulty in using the labs without training.

Most students find that the laboratories provide an enjoyable alternative to the "theorem-proof-example" format of a standard abstract algebra course. At the end of the semester, one student wrote in his evaluation of the course:

I am truly grateful for the laboratory component...Work on the computer helped to make the abstract theory more concrete... One of the best things about the labs was that we formed our own conjectures about the patterns we saw...I believe that the progression of (1) lab, (2) conjecture, (3) class discussion, and (4) proof was highly beneficial in gaining understanding of the abstract material of the course.

Table of Contents: 1. Groups and Geometry; 2. Cayley Tables; 3. Cyclic Groups and Cyclic Subgroups; 4. Subgroups and Subgroup Lattices; 5. The Center and Commutator Subgroups; 6. Quotient Groups; 7. Direct Products; 8. The Unitary Groups; 9. Composition Series; 10. Introduction to Endomorphisms; 11. The Inner Automorphisms of a Group; 12. The Kernel of an Endomorphism; 13. The Class Equation; 14. Conjugate Subgroups; 15. The Sylow Theorems; Appendix A. Table Generation Menu of *Exploring Small Groups (ESG)*; Appendix B. Sample Library of *ESG*; Appendix C. Group Library of *ESG*; Appendix D. Group Properties Menu

Exploring Small Groups, the software packaged with this lab manual, is on a 3½" DD PC compatible disk. This is a DOS program that can be run in Windows. The software was developed by Ladnor Geissinger, University of North Carolina at Chapel Hill.

112 pp., Paperbound, 1996
ISBN 0-88385-705-7
List: $22.00 MAA Member: $16.00
Catalog Code: LABEJR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
PO Box 91112, Washington, DC 20090-1112
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

Membership Code:

____ ____ ____ ____ ____ ____

Name ______________________

Address ______________________

City ______________________

State ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	LABE/JR	______	______
______	______________	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ______________ Expires ___/___

Signature ______________________

UNITED STATES POSTAL SERVICE™

Statement of Ownership, Management, and Circulation

(Required by 39 USC 3685)

1. Publication Title: The American Mathematical Monthly
2. Publication Number: 0 0 0 2 - 9 8 9 0
3. Filing Date: November 6, 1996
4. Issue Frequency: Monthly, except bimonthly June/July and August/Sept.
5. Number of Issues Published Annually: 10
6. Annual Subscription Price: $38
7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4): MAA, 1529 Eighteenth St., NW, Washington, DC 20036-1385

Contact Person: Harry Waldman
Telephone: 202-387-5200

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): same

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address): The Mathematical Association of America, 1529 Eighteenth St., NW, Washington, DC 20036-1385

Editor (Name and complete mailing address): Roger Horn, University of Utah, Salt Lake City, UT 84112

Managing Editor (Name and complete mailing address): Harry Waldman, MAA, 1529 18th St., NW, Washington, DC 20036

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name	Complete Mailing Address
The Mathematical Association of America	1529 Eighteenth St., NW, Washington, DC 20036

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at special rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, September 1995 (See Instructions on Reverse)

13. Publication Title: AMERICAN MATHEMATICAL MONTHLY
14. Issue Date for Circulation Data Below: NOVEMBER 1996

15. Extent and Nature of Circulation		Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)		18,750	18,275
b. Paid and/or Requested Circulation	(1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales (Not mailed)	0	0
	(2) Paid or Requested Mail Subscriptions (Include advertiser's proof copies and exchange copies)	17,500	16,344
c. Total Paid and/or Requested Circulation (Sum of 15b(1) and 15b(2))		17,500	16,344
d. Free Distribution by Mail (Samples, complimentary, and other free)		700	700
e. Free Distribution Outside the Mail (Carriers or other means)		0	0
f. Total Free Distribution (Sum of 15d and 15e)		700	700
g. Total Distribution (Sum of 15c and 15f)		18,200	17,044
h. Copies not Distributed	(1) Office Use, Leftovers, Spoiled	550	1,131
	(2) Returns from News Agents	0	0
i. Total (Sum of 15g, 15h(1), and 15h(2))		18,750	18,275
Percent Paid and/or Requested Circulation (15c / 15g x 100)		96%	96%

16. Publication of Statement of Ownership
☒ Publication required. Will be printed in the DECEMBER issue of this publication.
☐ Publication not required

17. Signature and Title of Editor, Publisher, Business Manager, or Owner: Donald J Albers, Dir. of Publication — Date: 11/6/96

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including multiple damages and civil penalties).

Instructions to Publishers

1. Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
2. In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation. In item 11, if none, check the box. Use blank sheets if more space is required.
3. Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
4. If the publication had second-class authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published; it must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
5. In item 16, indicate the date of the issue in which this Statement of Ownership will be published.
6. Item 17 must be signed.

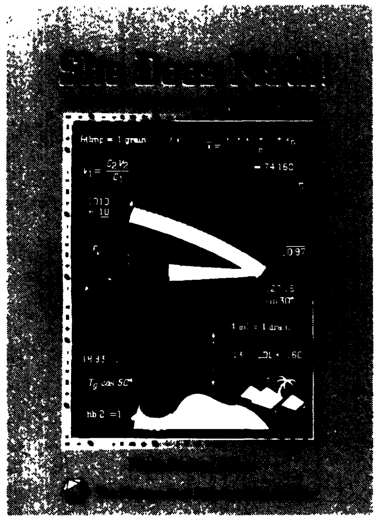
Failure to file or publish a statement of ownership may lead to suspension of second-class authorization.

PS Form 3526, September 1995 (Reverse)

ͻhe Does Math!

ƚeal-Life Problems From Vomen On The Job

.arla Parker, Editor



ıe Does Math! presents the career histories of 38 pro-ssional women and math problems related to their ork. Each history describes how much math the ıthor took in high school and college; how she chose :r field of study; and how she ended up in her current b. Each of the women presents problems that are typi-.l of those she has faced in her job. The problems quire only high school mathematics for their solu-ɔn.

ho should have this book?

ɔur daughter, your granddaughter, your sister, your rmer math teacher, your students—and young men o. They want to know how the math they study is ·plied, and this book will show them.

ıere are many reasons to buy this book:

By reading the career histories of the women profiled this book, young people will learn that if they take athematics courses in high school and college they ıll be qualified to enter interesting technical fields d earn good salaries.

The problems have special appeal to students who ɜ beginning to think about career choices.

The book provides practical information about the ɔ market in an interesting, innovative way.

Strong female role models who work as successful :hnical professionals are presented.

The problems are interesting and challenging, yet ʝuire only high school mathematics. They demonstrate w good math skills are applied to real-life problems.

ad what others have said about She Does Math!

ıally — a practical, innovative, well-written book ıt will also inspire its readers. The wonder is...it's a ıthematics text and a biography! The idea of ɔmen telling their own career stories, emphasizing the mathematics they use in their jobs is extremely creative. This book makes me wish that I could go through school all over again!

Anne L. Bryant, Executive Director
American Association of University Women

She Does Math! *will undoubtedly appeal both to students who already enjoy math and want to get a view of potential career paths, and also to students who want to better understand the relevance of their math classes to their future careers. It is an absorbing look into the lives of some very inspiring and talented women!*

Susanne Hupfer and Elisabeth Freeman
Yale University

This collection is a wonderful confirmation that real women do math. They do math in a surprising variety of careers, fully enjoying the challenge and rewards of solving complex problems. This is a book for young women and men, a book for their teachers and parents, a book that informs about the possibilities that mathematics affords to all. It is also a book that will engage you in real-life mathematics! – Doris Schattschneider
Moravian College

272 pp., Paperbound, 1995 ISBN 0-88385-702-2
List: $27.50 MAA Member: $20.95
Catalog Code: SDM/JR

ORDER FROM:
THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW Washington, DC 20036
1-800-331-1622 (301) 617-7800 FAX (301) 206-9789

:mbership Code: ___ ___ ___ ___ ___ ___

me ____________________

dress ____________________

:y ____________________

te ____________ Zip ____________

QTY.	CATALOG CODE	PRICE	AMOUNT
______	SDM/JR	______	______
______	______	______	______
		TOTAL	______

Payment ☐ Check ☐ VISA ☐ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ____________________